Annu. Rev. Biochem. 1992. 61:1053-95

TRANSCRIPTION FACTORS: Structural Families and Principles of DNA Recognition

Carl O. Pabo

Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Robert T. Sauer

Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

KEY WORDS: protein-DNA recognition, DNA-binding protein, helix-turn-helix, homeodomain, zinc finger

CONTENTS

0066-4154/92/0701-1053$02.00

INTRODUCTION

DNA-binding proteins play central roles in biology. Among other activities, they are responsible for replicating the genome, for transcribing active genes, and for repairing damaged DNA. One of the largest and most diverse classes of DNA-binding proteins are the transcription factors that regulate gene expression. In this review, we focus on structural studies of the DNA-binding domains from these transcription factors. A more general review of protein-nucleic acid interactions, which also includes a discussion of restriction enzymes, polymerases, and RNA-binding proteins, can be found in Steitz (1).

Transcription factors regulate cell development, differentiation, and cell growth by binding to a specific DNA site (or set of sites) and regulating gene expression. One cannot fully understand how genetic information is utilized without understanding the structure and DNA-binding properties of these transcription factors. In this review, we address progress in understanding the structural basis for sequence-specific binding. What secondary structures can provide a surface that is complementary to the structure of double-helical DNA? What contacts with the bases and the DNA backbone allow site-specific recognition? How does understanding these structural details enhance our understanding of the molecular mechanisms involved in repression and activation of gene expression?

There have been several recent reviews of transcription factors and DNA-binding proteins (1–7). In addition, there have been reviews focusing on specific families of transcription factors such as the helix-turn-helix proteins (8–10); zinc finger, steroid receptor, and other metal-binding DNA-binding proteins (11–17); leucine zipper proteins (18); homeodomains (19–22); and β-sheet DNA-binding proteins (23). There also have been separate reviews covering the Trp repressor (24) and λ repressor (25). When we wrote our last general review (26), structures were known for only three DNA-binding proteins. Now structures have been reported for more than 10 protein-DNA complexes and for more than 20 DNA-binding proteins. This wealth of new data allows a much broader perspective on the general problem of protein-DNA recognition. Since an exhaustive review of the literature would be overwhelming, we have chosen to focus on a few well-studied systems that illustrate the main biological and structural issues.

FAMILIES OF DNA-BINDING PROTEINS

One of the central observations emerging from structural studies and sequence comparisons is that many DNA-binding proteins can be grouped into classes that use related structural motifs for recognition. Some families, such as the helix-turn-helix proteins, were first recognized because of structural similarities. Other families were first identified by sequence comparisons and later characterized by structural studies. Large, well-established families include the helix-turn-helix (HTH) proteins, the homeodomains, zinc finger proteins, the steroid receptors, leucine zipper proteins, and the helix-loop-helix proteins. Two smaller families have been identified that use β-sheets for DNA binding. Sequence comparisons indicate that there are a number of additional families of DNA-binding proteins, but fewer structural data are available for the families that have been characterized most recently.

Familial relationships are a powerful unifying theme in the study of DNA-binding proteins since they relate evolution, structure, recognition, gene regulation, and design. Thinking about these motifs can—at least in a schematic fashion—give a satisfying general picture of protein-DNA recognition. First, the existence of distinct families shows that there are multiple solutions to the structural problem of designing DNA-binding proteins. There is no single pattern or simple code—evolution has solved the problem in several different ways. However, each of the motifs that has been characterized in detail involves a simple secondary structure (usually an α-helix) that is complementary to the structure of B-DNA in a straightforward way. Side chain contacts play a major role in site-specific binding and often allow the same motif to be used at a set of different sites. In some way, the number of members in a particular family of DNA-binding proteins may measure the relative "evolutionary success" of a particular DNA-binding motif. These major families are nature's most successful designs for DNA-binding proteins.

Although thinking about DNA-binding proteins in terms of families has many attractive aspects, it is important to realize that there are many interesting and important DNA-binding proteins that do not belong to any of the known families. The SV40 large T antigen (27) and the human p53 tumor suppressor gene (28) are two such examples. Of course, there is no reason to believe that we have discovered all the major families of DNA-binding proteins, or to believe that all DNA-binding proteins must belong to clearly identified families. In spite of these cautions, it seems worthwhile to focus on the major families of DNA-binding proteins because information about a single member provides a framework for thinking about the whole family. In addition, comparisons within families and between families of DNA-binding proteins may help us understand the general principles of protein-DNA recognition. Finally, since these motifs have been so successful in proliferating and adopting new roles during evolution, it seems reasonable to imagine that they may also provide the most convenient scaffolds to use in attempting to design new DNA-binding proteins.

Helix-Turn-Helix

The helix-turn-helix (HTH) structure was the first DNA-recognition motif discovered, and structures have now been determined for many HTH proteins

and protein-DNA complexes. Since the HTH proteins have been analyzed in such detail, studying these proteins provides an excellent background for thinking about other motifs that are used in site-specific recognition. The first crystal structures of DNA-binding proteins were those of the λ Cro protein (29, 30), the *Escherichia coli* CAP protein (31), and the DNA-binding domain of λ repressor (32). Comparisons of these three proteins revealed a conserved recognition motif consisting of an α-helix, a turn, and a second α-helix (33, 34). Structures are now available for several other HTH proteins, including Lac repressor (35), 434 repressor (36, 37), 434 Cro (38, 39), Trp repressor (40), and Fis protein (41). The DNA-binding domain of LexA contains a related two-helix motif, but the turn region is more extended than in the other HTH proteins (42).

Although structural studies first led to its identification, sequence comparisons suggested that the HTH motif occurred in a large family of prokaryotic DNA-binding proteins (43–45). The most highly conserved residues in the HTH motif include a glycine in the turn and also include several hydrophobic residues (see Figure 1). These conserved residues help to stabilize the arrangement of the two helices in the HTH unit and help this unit to pack against the rest of the protein.

It is worth emphasizing that the HTH motif (unlike many of the other motifs that we discuss) is not a separate, stable domain. The HTH motif cannot fold or function by itself but always occurs as part of a larger

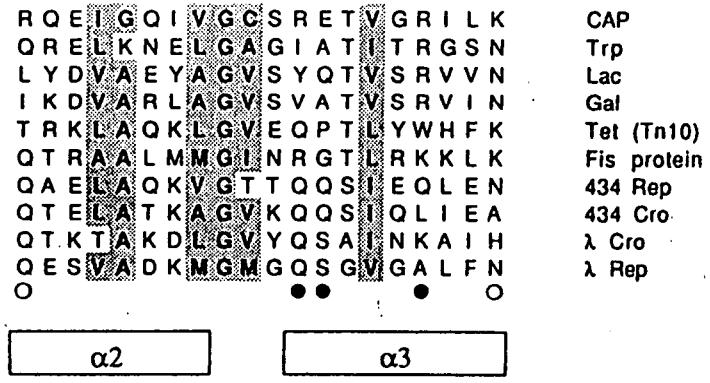

Figure 1 Alignment of the helix-turn-helix sequences of phage and bacterial regulatory proteins. Sequence positions that are generally hydrophobic (A, C, V, I, L, M, F, Y) or small (G, A) are shaded. The open and closed circles represent residue positions that make backbone contacts (○) or that mediate hydrogen bonds with bases in the major groove (●) in the λ and/or 434 repressor-complexes (47, 48). For sequence references, see (26) and references therein and (46).

DNA-binding domain. Comparing the first HTH protein structures immediately revealed that the HTH motif occurred in a number of different structural environments. Thus, the λ Cro protein and the *E. coli* CAP protein contain a number of β-sheets in their DNA-binding domains, while the corresponding domain of λ repressor is entirely α-helical. Studies of complexes involving HTH proteins (discussed below) have revealed that other regions of the DNA-binding domains (outside of the HTH units) can also have significant roles in recognition. Although there is a natural tendency to focus on contacts made by the HTH unit, it is wrong to assume that recognition involves only such contacts. It is even more misleading to focus exclusively on the contacts made by the second helix—often called the "recognition" helix—of the HTH motif.

The structures of several repressor-operator complexes have been determined at high resolution, yielding a wealth of information about how the HTH motif is used in site-specific recognition. The structure of the λ repressor-operator complex (47) is used to illustrate our discussion (Figure 2). The N-terminal domain of λ repressor forms a dimer, and each subunit interacts with one-half of the operator site. In each half-site, side chains from helices 2 and 3 of λ repressor (the HTH unit) make critical contacts with the DNA. For example, Gln44, the first residue of helix 3, makes two hydrogen bonds with an adenine near the outer edge of the operator; Ser45 makes a single hydrogen bond with a guanine; Gly46 and Gly48 make hydrophobic contacts with thymine methyl groups; and Asn52 makes a hydrogen bond to a phosphodiester oxygen. Many of these contacts are shown in Figures 2*B* and 2*C*. Gln33, in helix 2, contacts a phosphodiester oxygen and also hydrogen bonds to the Gln44 side chain. Presumably, this side chain–side chain interaction (Figure 2*C*), which illustrates the structural complexity of the recognition process, helps to stabilize both the backbone contact made by Gln33 and the base contacts made by Gln44. It is noteworthy that 434 repressor has a pair of glutamines at corresponding positions, which have similar roles in recognition of the 434 operator (48, 49).

As mentioned above, residues outside of the HTH unit also are important for DNA recognition. In the λ complex, Lys19 and Tyr22 illustrate how residues from neighboring regions can make contacts with the DNA backbone (Figure 3). λ repressor also has a distinct structural motif—an extended N-terminal arm—that makes critical contacts in the major groove. Lys4, in the N-terminal arm, cooperates with Asn55 (in the loop region following helix 3) to make two hydrogen bonds with a guanine in the major groove (Figure 2*C*). The rest of the N-terminal arm wraps around the center of the operator and makes several additional contacts (47, 50). The N-terminal arms of λ repressor are flexible in solution (32, 51, 52) and only assume a fixed position upon operator binding (47, 50). As we will see, similar disorder → order

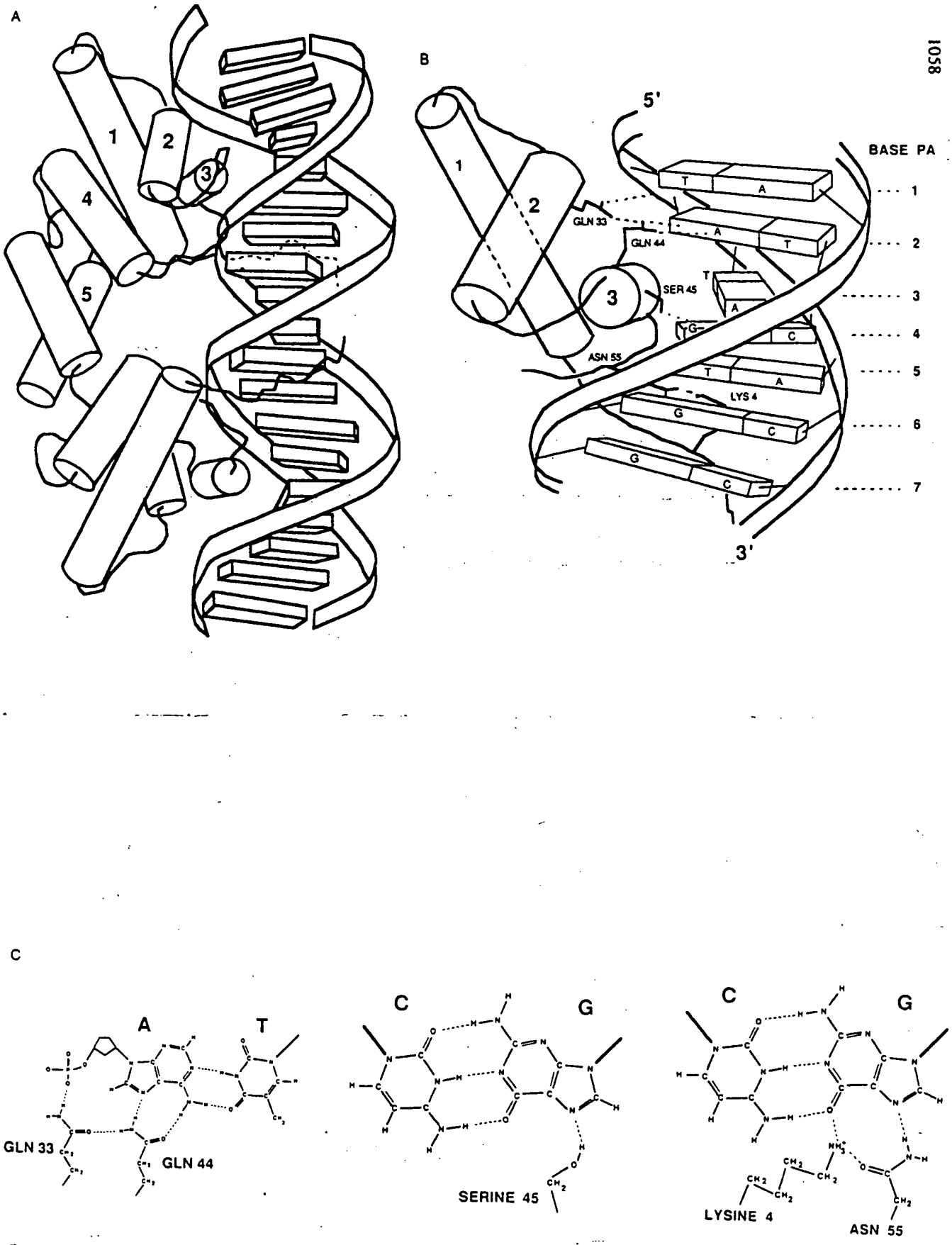

Figure 2 *A*. Sketch showing overall arrangement of complex containing the N-terminal domain of λ repressor (47). Cylinders are used to represent α-helices, and the helices in one monomer are numbered. Helices 2 and 3 correspond to the conserved HTH unit. *B*. Sketch emphasizing the HTH unit and hydrogen bonds with the bases in one-half of the λ operator site. *C*. Sketches showing hydrogen bonds with bases 2, 4, and 6 in one-half of the λ operator site. Figures reprinted with permission from *Science*.

transitions occur in other DNA-binding motifs (such as the basic region of the leucine zipper proteins).

Crystal structures of complexes have also been reported for the 434 repressor (48), 434 Cro (38, 53), Trp repressor (54), λ Cro (55), and *E. coli* CAP protein (56). These cocrystal structures show that all of the HTH protein-DNA complexes have a number of common features: (*a*) The repressors bind

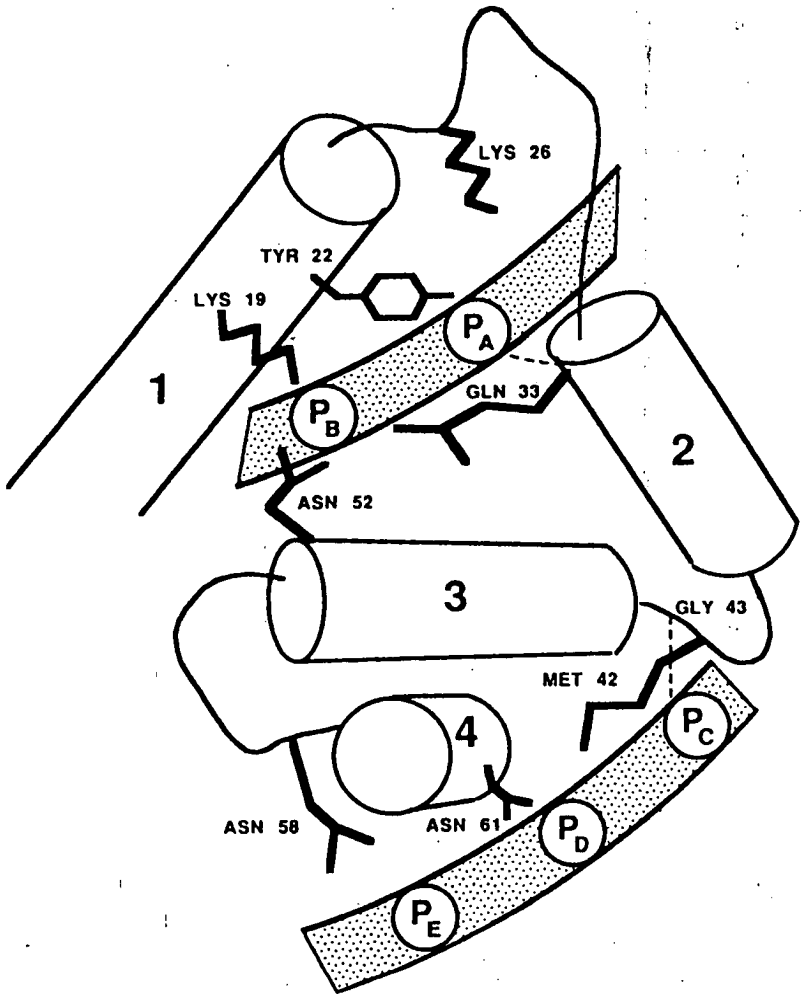

Figure 3 Sketch showing side chains that interact with the sugar-phosphate backbone in one-half of the λ repressor-operator complex (47). Reprinted with permission from *Science*.

as dimers. Each monomer recognizes a half-site, and the approximate symmetry of the DNA-binding site is reflected in the approximate symmetry of the complexes. (*b*) The conserved HTH unit contacts the DNA in each half of the operator site. The first helix of the HTH unit is somewhat "above" the major groove, but the N-terminus of this helix contacts the DNA backbone. The second helix of the HTH unit fits into the major groove, and the N-terminal portion of this helix is closest to the edges of the base pairs.[1] (*c*) The operator sites are B-form DNA. (*d*) Side chains from the HTH units make site-specific contacts with groups in the major groove. (*e*) Each complex has an extensive network of hydrogen bonds between the protein and the DNA backbone. Some of these contacts are made by lysines or arginines, but many of them are made by short, polar side chains or even by -NH groups from the polypeptide backbone.

Since a set of high-resolution structures is available, the HTH complexes can be used to illustrate principles that may be used by other families of DNA-binding proteins. As mentioned above, λ repressor uses an extended region of peptide chain to wrap around the DNA and augment the contacts made by residues in the HTH region. This reveals another motif that may be important in protein-DNA recognition and illustrates how several distinct structural motifs can contribute to site-specific binding. The 434 repressor is sensitive to base substitutions in the center of its operator site, even though it does not directly contact any of these bases (36, 48). The central bases in the 434 operator are thought to influence binding indirectly via effects on DNA conformation (57, 58). Similar effects may play a role in protein-DNA recognition in a variety of systems. The Trp repressor appears to use several water molecules to provide critical contacts (54). These waters, which are tightly bound at the N-terminal end of an α-helix, show that hydration can have important effects on recognition, and further illustrate the structural complexity of protein-DNA interactions.

Although we have focused on the DNA-binding domains, the HTH proteins often contain additional domains that have important roles in regulating activity. For example, the N-terminal domain of CAP allows dimer formation and also binds cAMP, an allosteric effector of DNA binding (31). In the λ, 434, and LexA repressors, the C-terminal domains allow stable dimer formation, and the process of induction involves proteolytic cleavage between the N-terminal and C-terminal domains (for review, see 25). This modular arrangement, with different functions in different domains, also is a common theme in eukaryotic transcription factors, where the DNA-binding domains represent only part of the intact proteins (2–6).

[1]Although each of these proteins uses the HTH unit in a generally similar way, there are some interesting differences in the precise arrangement of the HTH unit with respect to the DNA [for review, see Harrison & Aggarwal (9)].

Homeodomain

The homeodomain is a DNA-binding motif that is present in a large family of eukaryotic regulatory proteins (19, 20). Although the conserved sequences were first recognized in proteins that regulate *Drosophila* development (see Figure 4), it now is clear that the homeodomain has a broader role in eukaryotic gene regulation. A comparison of amino acid sequences suggested that the homeodomain would contain a HTH motif (59, 60), and this has now been confirmed by structural studies. However, unlike the isolated HTH unit, the 60-residue homeodomain forms a stable, folded structure and can bind DNA by itself (61–63).

The structure of the homeodomain was first determined by 2D NMR studies of the *Drosophila* Antennapedia (Antp) homeodomain (63, 64), and there have now been several studies of homeodomains (64a) and homeodomain-DNA complexes (65–67). These studies confirmed that the homeodomain contains a HTH motif. Distance constraints, determined from 2D NMR studies of the Antp complex, were used to dock the homeodomain against a B-form DNA site (65). The crystal structure of the *Drosophila* engrailed homeodomain-DNA complex has also been solved (66), and the structure of a complex containing the yeast MAT α2 homeodomain has recently been determined (67).

The engrailed homeodomain contains an extended N-terminal arm and three α-helices (Figure 5). The overall structure of the protein is quite easy to visualize: helix 1 and helix 2 pack against each other in an antiparallel arrangement. Helix 3 is roughly perpendicular to the first two helices, and the hydrophobic face of this extended helix packs against helices 1 and 2 to form the interior of the protein. The C-terminal residues of the isolated Antp homeodomain have been described as a fourth helix, separated from helix 3 by a kink (63). However, in the engrailed cocrystal and α2 cocrystal these C-terminal residues form a continuous, helical extension of helix 3 (66, 67).

The main contacts in the engrailed complex are made by residues in helix 3, which fits into the major groove, and by residues in the extended N-terminal arm, which fits into the minor groove (Figures 5 and 6). Most of the contacts are made by helix 3. The exposed face of helix 3 fits directly into the major groove, allowing side chains to make extensive contacts with the bases and with phosphodiester oxygens along the edge of the major groove. Residues near the middle of helix 3 are closest to the bases, but residues near the C-terminal end make a number of contacts with the DNA backbone (Figures 4–6). Helices 1 and 2 span the major groove but are much farther from the DNA. There only are two DNA contacts made by this region of the protein. Both are with the DNA backbone. The other critical contacts come from an N-terminal arm that, as predicted by Garcia-Blanco et al (68), has some similarities with λ repressor's arm. Residues 3–9 of engrailed form an ex-

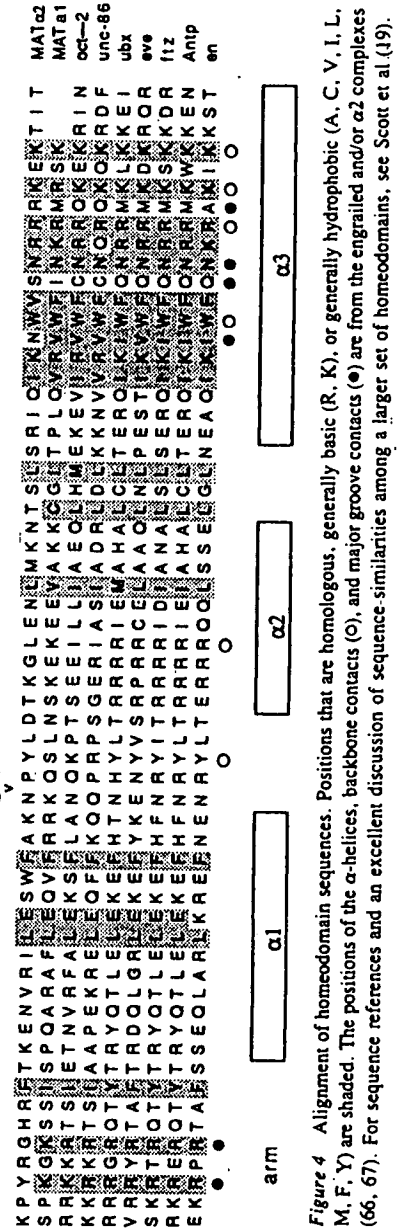

Figure 4 Alignment of homeodomain sequences. Positions that are homologous, generally basic (R, K), or generally hydrophobic (A, C, V, I, L, M, F, Y) are shaded. The positions of the α-helices, backbone contacts (O), and major groove contacts (●) are from the engrailed and/or α2 complexes (66, 67). For sequence references and an excellent discussion of sequence-similarities among a larger set of homeodomains, see Scott et al (19).

tended N-terminal arm that fits into the minor groove and supplements the contacts made by helix 3. (Since the homeodomain is just a small part of the intact engrailed protein, this extended N-terminal arm probably would form a loop or linker region in the intact protein.)

The crystal structure of the α2 homeodomain-DNA complex has recently been determined (67), and turns out to be very similar to the engrailed complex. This is particularly interesting because the α2 sequence is one of the most divergent within the homeodomain family. (Some scientists had refused to recognize α2 as a legitimate member of the family.) However, α2 and engrailed fold in very similar ways: the C_α backbone of α2 superimposes quite well on the C_α backbone of engrailed (67). The α2 homeodomain has a three-residue "insertion" in the loop between helices 1 and 2, but this is accommodated without any major changes in the overall fold of the protein. Since α2 differs substantially from engrailed in sequence but has a similar structure, it will be interesting to see whether members of different homeodomain subfamilies, which have been defined by grouping proteins with the most similar sequences (19), have any distinctive structural or functional properties.

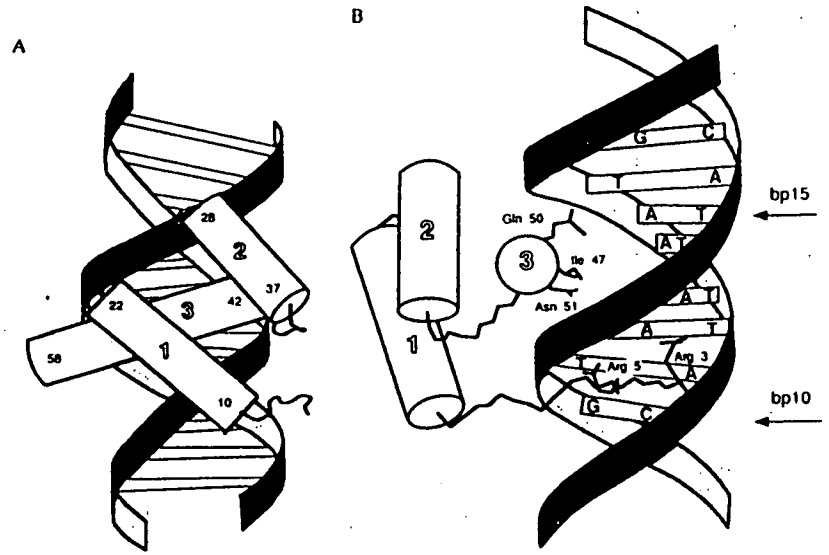

Figure 5 *A*. Sketch of the engrailed homeodomain-DNA complex (66), summarizing the overall relationship of the α-helices and the N-terminal arm with respect to the DNA. Helices 2 and 3 form the conserved HTH unit. *B*. View, at right angles to that shown in panel A, emphasizes the minor groove contacts made by the N-terminal arm and the major groove contacts made by helix 3. Reprinted with permission from *Cell*.

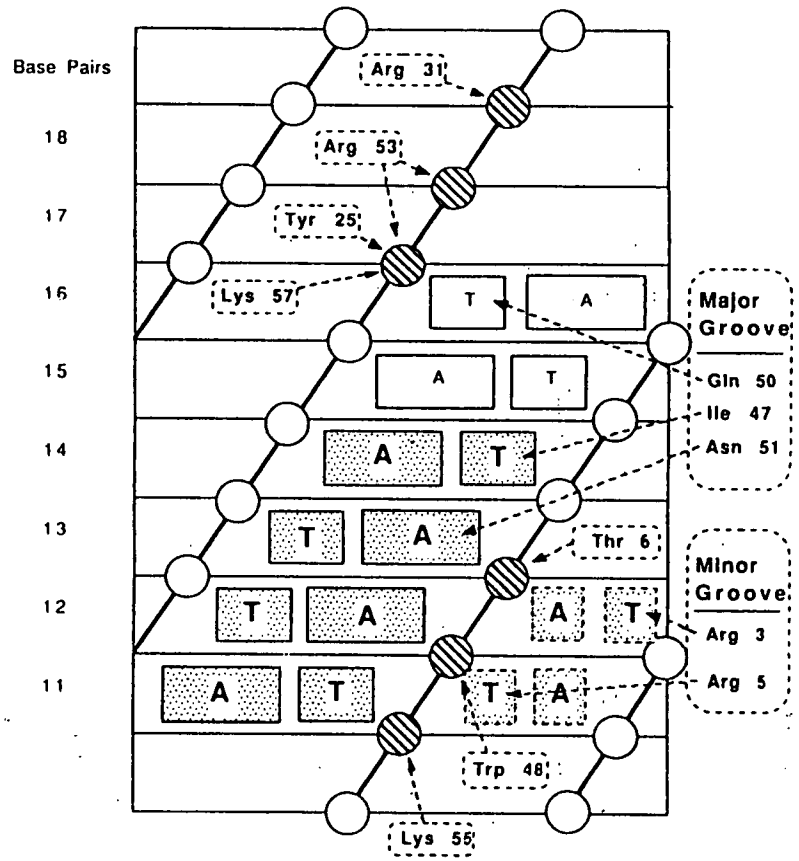

Figure 6 Sketch summarizing all the contacts in the engrailed homeodomain-DNA complex (66). The DNA is represented as a cylindrical projection, and the shading emphasizes the TAAT subsite that occurs in many homeodomain-binding sites. Phosphates are represented by circles; hatched circles show phosphates that are contacted by the engrailed homeodomain. Reprinted with permission from *Cell*.

A comparison of the engrailed and α2 complexes also reveals that these two homeodomains dock against the DNA in very similar ways (67). As observed with the engrailed complex (66) and with the Antp complex (65), the N-terminal arm makes contacts in the minor groove, and helix 3 makes an extensive set of contacts in the major groove. Superimposing the engrailed and α2 complexes by superimposing helix 3 of each protein shows that there are a set of common contacts, which are mediated by conserved residues at the protein-DNA interface. There are two conserved hydrogen bonds with an

adenine in the major groove. These are made by Asn51, which is one of the four "invariant" residues found in every homeodomain (19). Six other conserved residues make conserved contacts with phosphodiester oxygens that flank the major groove. Presumably, these sets of conserved contacts play a critical role in determining the precise orientation of helix 3.

Since the $\alpha 2$ and engrailed complexes are so similar, they may provide a basis for modeling other homeodomain-DNA complexes. Further studies will be needed before we can fully explain the different sequence preferences of the homeodomains, but several residues are appropriately positioned to make base contacts and are likely to influence DNA specificity. Residues 47, 50, 51, and 54 appear to be especially important. Genetic and biochemical analyses have shown that residue 50 is important for controlling the differential specificity of homeodomain-DNA interactions (69–71), and the structures show that the side chain of residue 50 points into the major groove and is in an excellent position to contribute to the specificity of binding (Figure 5*B*). The structures also show that that the side chains of residues 47 and 54 can make base contacts (although the Ala54 side chain of engrailed is too short to reach the bases). As noted above, Asn51 hydrogen bonds to an adenine in each complex. Presumably, this interaction is important for binding, but Asn51 cannot determine differential specificity since it is conserved throughout the homeodomain superfamily. Differences in the sequence and in the precise orientation of the N-terminal arm may also help determine binding specificity. Since engrailed and $\alpha 2$ use slightly different regions of the arm, we will need more structures and more biochemical data before we can understand how these contacts contribute to specificity.

Overall, the structural data agree very well with biochemical and genetic data about the homeodomains and about homeodomain-DNA interactions. For example, Trp48, Phe49, Asn51, and Arg53 occur in every one of the higher eukaryotic homeodomains compiled by Scott et al (19). These invariant residues occur right in the middle of helix 3, in the region that is closest to the major groove. Trp48 and Phe49 form part of the hydrophobic core and must play a key role in stabilizing the correct folded structure (Trp48 also makes a backbone contact). The invariant hydrophilic residues—Asn51 and Arg53—make critical contacts with the DNA. Asn51 makes a pair of hydrogen bonds with an adenine (Figure 7), and Arg53 hydrogen bonds with two phosphate groups on the DNA backbone (Figure 6). The structure also explains the roles of other highly conserved residues and regions of the homeodomain, including residues in the N-terminal arm, other residues in the hydrophobic core, and the set of residues that contact the phosphodiester backbone.

As first predicted from sequence comparisons (59, 60), the homeodomain contains a HTH unit (63) that is similar to those observed in the prokaryotic

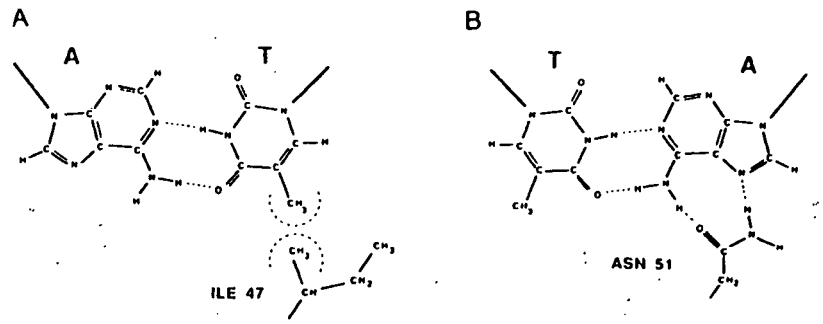

Figure 7 Sketch showing the major groove contacts made by *A*. Ile47 and *B*. Asn51 in the engrailed homeodomain-DNA complex (66). Reprinted with permission from *Cell*.

repressors. However, structural studies of the complexes (65–67) have shown that the homeodomain uses the HTH unit in a novel way. These differences can be illustrated by comparing the λ repressor-operator and engrailed homeodomain-DNA complexes (66). Superimposing the λ and engrailed HTH units (to provide a common frame of reference) leaves the DNA sites in very different positions, showing that these HTH units dock against the DNA in significantly different ways. In the λ complex, the N-terminal portion of helix 3 (the "recognition helix") is closest to the bases. In the engrailed complex, the center of a much longer helix 3 (again, the "recognition helix") is closest to the bases. Given these differences, it is not surprising that the critical base contacts are made by residues from different portions of the HTH units. In the λ and 434 complexes, the critical base contacts are made by residues in the first turn (i.e. at the N-terminal end) of the recognition helix. In the homeodomain complexes, residues in the second and third turns of the recognition helix make critical contacts with the bases. There also are significant differences in the way that helix 2 (the first helix of the HTH unit) is used in the two complexes. In the λ and 434 complexes, the N-terminal end of helix 2 fits partway into the major groove and the peptide -NH of the first residue hydrogen bonds to a phosphodiester oxygen on the DNA backbone. In contrast, helix 2 of the homeodomain lies entirely above the major groove and cannot use a peptide -NH to hydrogen bond to the DNA.

The different docking arrangements seen for the prokaryotic and eukaryotic HTH units pose some challenging questions about the the evolution and structural significance of the HTH unit. However, it is important to recognize that the helices in the HTH unit of the homeodomain are significantly longer than the corresponding helices in many of the prokaryotic HTH proteins.[2] These differences in helix length presumably are responsible for some of the differences in the docking arrangement. Thus lengthening the first helix of the HTH unit probably prevents this helix from tucking partway into the major

groove. To avoid van der Waals collisions with the DNA backbone, this extended helix must lie entirely above the major groove, and this requires that the HTH unit dock against the DNA in a significantly different way. The C-terminal extension of helix 3 may also be critical in stabilizing the docking arrangement seen with the homeodomain HTH motif. In engrailed and α2, three residues from this C-terminal extension (Arg53, Lys55, and Lys57) contact the DNA backbone and help to hold the helix in the major groove (Figure 6). When comparing HTH motifs and docking arrangements, it is important to remember that the prokaryotic proteins present the HTH unit in a variety of structural contexts. Future studies may yet reveal prokaryotic HTH motifs that dock as the homeodomain does (72) or may reveal other docking arrangements for the homeodomain, and thus provide a "missing link" that helps us understand the evolutionary and structural relationships between these families.

Although an isolated homeodomain can fold correctly and bind DNA with a specificity similar to that of the intact protein, it seems likely that the precise DNA-binding specificity is modulated by other regions of the protein. Many homeodomain proteins contain other sequence motifs that flank the homeodomain and are conserved within specific subfamilies (19). As noted by Kissinger et al (66), both the N-terminus and C-terminus of the homeodomain are near the DNA, and neighboring residues from the intact protein could easily contact flanking regions of the DNA. The POU proteins (73, 74) seem to provide a particularly clear example of how this may occur. The POU-specific domain, which was first observed in the pit-1, oct-2, and unc-86 proteins, is a conserved 65–75-residue segment just on the N-terminal side of the homeodomain (75). This POU-specific domain appears to make DNA contacts with a set of bases adjacent to those contacted by the homeodomain (73, 74). It is possible that other conserved sequences found in the intact homeodomain proteins have other important roles in DNA recognition. They might influence contacts made by the N-terminal arm or the C-terminal helix, and they could provide "attachment" or "targeting" sites for other proteins that modulate the specificity or affinity of DNA binding.

Protein-protein interactions may also have a role in modulating many homeodomain-DNA interactions. For example, the yeast α2 protein forms homodimers, but also forms complexes with a related homeodomain protein a1 and with the general transcription factor Mcm1 (76, 77). Each of these complexes has different site preferences, and it is possible that these accessory proteins (in addition to adding new DNA contacts) actually alter alter the way that α2 interacts with DNA. Recent studies have shown that mutations in the α2 HTH motif have different effects on the binding of α2 and on the binding of the a1/α2 complex (A. Vershon and A. Johnson, unpublished). Studies of the human oct-1 homeodomain protein have also emphasized the importance of accessory proteins in recognition and regulation (78), but additional studies will be needed to understand precisely how these proteins influence recognition. Do they provide additional contacts, or do they affect the way that the homeodomain docks against the DNA? Can they change the orientation of the homeodomain recognition helix or shift the position of the N-terminal arm?

[2]It is interesting to consider how our views might be influenced by the fact that the bacterial HTH proteins were solved first. Had the homeodomains been solved first, their extended helices would have been used to define the HTH unit, and subsequent structural comparisons with the prokaryotic HTH proteins would have yielded a far less impressive degree of structural homology.

Zinc Finger

Zinc fingers, of the type first discovered in the *Xenopus* transcription factor IIIA (TFIIIA) (79, 80), are another one of the major structural motifs involved in protein-DNA interactions (11–14). Zinc finger proteins are involved in many aspects of eukaryotic gene regulation. Homologous zinc fingers occur in proteins induced by differentiation and growth signals, in proto-oncogenes, in general transcription factors, in genes that regulate *Drosophila* development, and in regulatory genes of eukaryotic organisms (11–14 and references therein). Proteins in this family usually contain tandem repeats of the 30-residue zinc finger motif, with each motif containing the sequence pattern Cys-X_{2or4}-Cys-X_{12}-His-X_{3-5}-His (Figure 8). Unfortunately, the term "zinc finger" has acquired a loose—almost topological—definition and has been used when referring to almost any sequence that has a set of cysteines and/or histidines within a short region of polypeptide chain. Here, we focus on fingers that are structurally homologous to the fingers in TFIIIA. Other cysteine-rich and histidine-rich motifs, such as those that occur in the steroid receptors, in the yeast transcription factor GAL4, and in certain retroviral proteins, are discussed later in this review.

Model building predicted (84, 85) and 2D NMR studies confirmed (86, 87) that the TFIIIA-like zinc fingers contain an antiparallel β-sheet and an α-helix. Two cysteines, which are near the turn in the β-sheet region, and two histidines, which are in the α-helix, coordinate a central zinc ion and hold these secondary structures together to form a compact globular domain. The crystal structure of a zinc finger–DNA (83) complex containing three fingers from zif268 (88) and a consensus zif-binding site has been reported. The crystal structure shows that the zinc fingers bind in the major groove of B-DNA and wrap partway around the double helix (Figure 9). Each finger has a similar way of docking against the DNA and makes base contacts with a three-base-pair subsite. [Sequence analyses and mutational studies had correctly predicted many features of this complex (89).] Neighboring fingers are

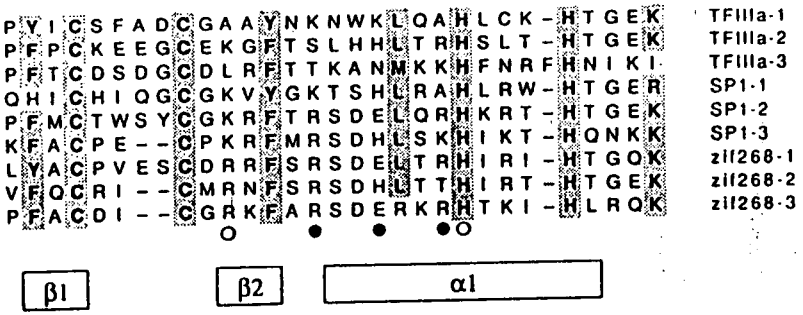

Figure 8 Alignment of zinc finger sequences. Positions of the conserved cysteines (C) and histidines (H) are shaded, as are positions that are generally basic (R, K) or hydrophobic (A, C, V, I, L, M, F, Y). The positions of secondary structure, conserved backbone contacts (o), and major groove contacts (•) are from finger 1, 2, and/or 3 of the zif268 complex. For sequence references see (81, 83, 88, 90). The sequence alignment is based upon that of Miller et al (79). An excellent discussion of sequence relationships between a much larger set of zinc finger proteins can be found in Krizek et al (82).

arranged in a way that reflects the helical pitch of the DNA and the three-base-pair periodicity of the subsites. Thus, a rotation of approximately 96 degrees (3 × 32 degrees/bp) around the DNA axis and a translation of approximately 10 Å (3 × 3.4 Å/bp) along the DNA axis move one finger onto the next. In the zif complex, the antiparallel β-sheet is on the back of the α-helix—away from the base pairs—and the β-sheet is shifted towards one side of the major groove. The first β-strand does not make any contacts with the DNA, whereas the second β-strand contacts the sugar phosphate backbone.

The zif268 fingers make a set of hydrogen bonds with bases in the major groove (see Figures 10 and 11). In each finger, critical base contacts are made by an arginine that immediately precedes the α-helix. This arginine also interacts with an aspartic acid that is the second residue in the α-helix. In finger 2, a histidine (the third residue in this helix) makes an additional base contact. In fingers 1 and 3, an arginine (the sixth residue in these helices) also makes a base contact. All of these contacts involve hydrogen bonds with guanines on the G-rich strand of the consensus binding site (5'-GCGTGGGCG-3'). The peptide binds with finger 1 near the 3' end (and with finger 3 near the 5' end) of this primary strand. The structure also reveals a set of contacts with the DNA backbone, and most of these also involve the primary, guanine-rich, strand of the DNA. In each finger, an arginine in the second β-strand makes a contact to a phosphodiester oxygen, and the first zinc-binding histidine in the α-helix contacts another phosphodiester oxygen (these contacts are indicated schematically by open circles in Figure 8).

The conserved arrangement of the fingers, and the pattern of side chain–

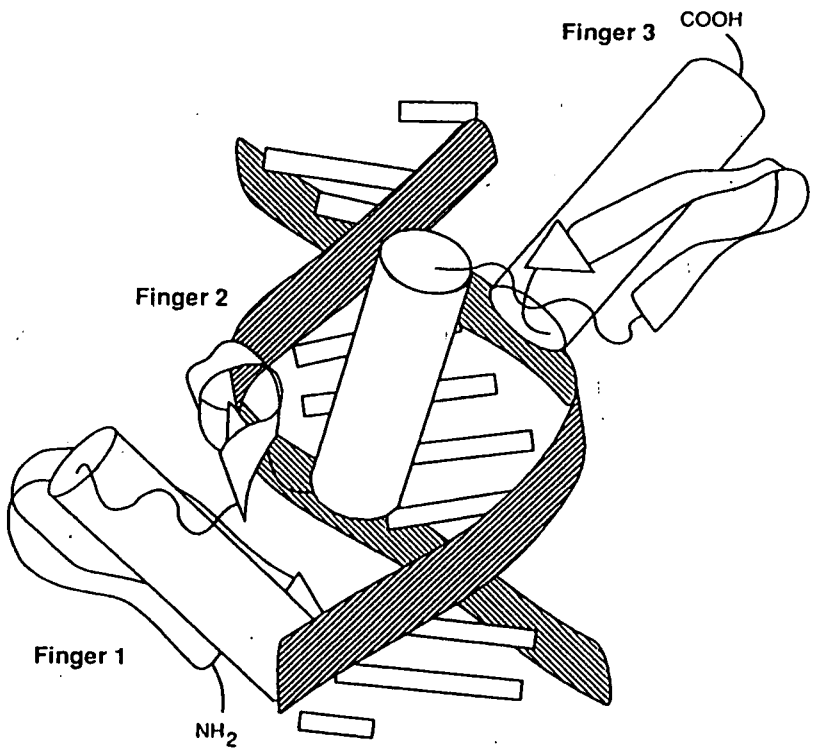

Figure 9 Sketch showing the overall arrangement of the zif268 zinc finger–DNA complex (83). The α-helices are shown as cylinders; β-sheets are shown as ribbons; and the zinc ions have been omitted from this sketch. The three fingers fit in the major groove, and each has a similar relationship to the DNA. Each finger makes base contacts in a three-base-pair subsite. Reprinted with permission from *Science*.

base interactions, are so regular that it may eventually be possible to describe a "code" for zinc finger–DNA interactions. The zinc finger uses the arginine that immediately precedes the α-helix, as well as the second, third, and sixth residues of the α-helix to contact the base pairs. (In the zif complex, the second residue of each helix is an aspartic acid, but a longer side chain at this position might be able to contact the bases.) Overall, there is a relatively simple pattern to the contacts (83). None of the zif fingers contacts all three bases, but the residue immediately preceding the α-helix contacts the third base on the primary strand of the subsite (5' — — G), the third residue in the α-helix can contact the second base on the primary strand (5' — G —), and

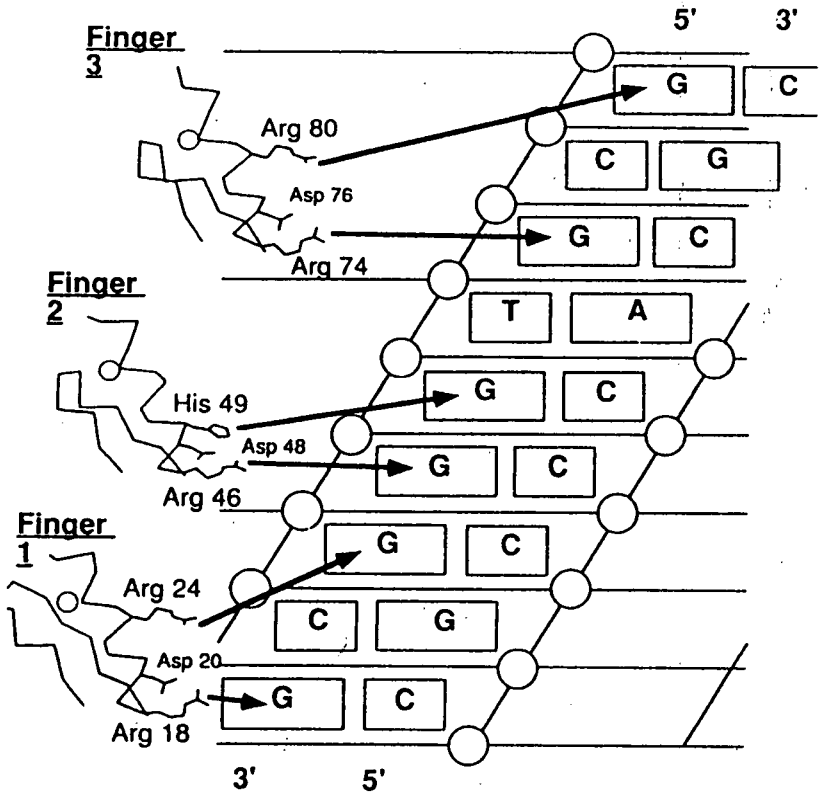

Figure 10 Sketch summarizing base contacts in the zif268 zinc finger–DNA complex (83). The DNA is represented as a cylindrical projection, and arrows indicate contacts in the major groove. The base contacts involve the guanine-rich strand of the binding site. Reprinted with permission from *Science*.

the sixth residue in the α-helix can contact the first base (5' G — —) of the subsite. These simple patterns reflect the fact that each of the three fingers docks against the DNA in a very similar way and is related to the next by a simple helical motion. This appears to be the first instance where the periodicity of a protein structure has such a simple relationship to the periodicity of double-helical DNA. Recognition is based on a simple modular system that can be used to recognize extended, asymmetric sites.

The structure of the zif268 complex should provide a useful guide for modeling complexes, such as Sp1 (90), with closely related fingers. However, there is no reason to believe that all fingers bind in exactly the same way. As we have discussed, the structurally conserved HTH unit can dock against

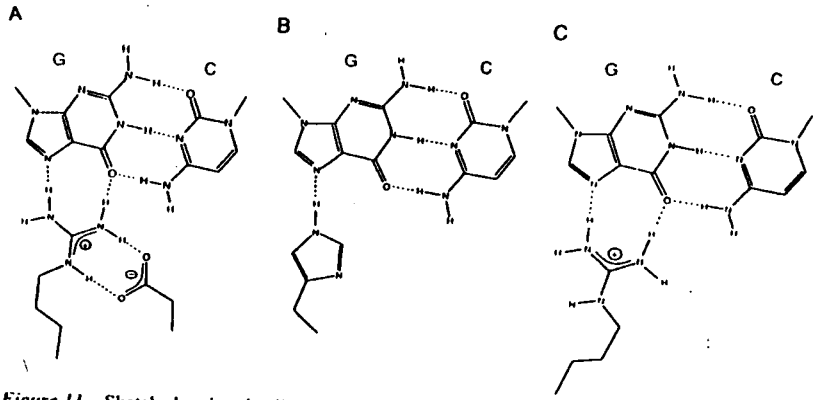

Figure 11 Sketch showing details of the base contacts in the zif268 zinc finger–DNA complex (83). *A*. Conserved contacts made by Arg18, Arg46, and Arg74. *B*. Contact made by His49 in finger 2. *C*. Conserved contacts made by Arg24 in finger 1 and Arg80 in finger 3. Reprinted with permission from *Science*.

the DNA in a number of different ways, and it seems plausible that the zinc finger motif may also have a set of different binding modes. TFIIIA-like zinc fingers are extremely common, and it seems that evolution has had ample opportunity to try different ways of using these fingers. Subtle sequence variants exist, such as the patterns found in the alternating fingers of ZFY (91), and these distinctive sequence patterns may reflect differences in the structure or docking arrangements of the fingers (92). Other distinctive sequence patterns have been noted. For instance, whenever there are five residues between the histidines, this finger usually is the last one in a tandem array (93) or occurs as a separated finger (94, 95). It will be necessary to solve the structures of other zinc finger–DNA complexes to see if other stable docking arrangements exist, to determine whether zinc fingers bind to AT-rich sequences in fundamentally different ways than they bind to GC-rich sequences, and to understand how zinc finger proteins interact with RNA [as in the 7S particle formed by TFIIIA (79 and references therein)].

Steroid Receptor

The steroid receptors are an important family of regulatory proteins that include receptors for the steroid hormones, retinoids, vitamin D, thyroid hormones, and a number of other important compounds. Genetic and biochemical studies done in a number of different laboratories revealed that these proteins contain separate domains for hormone binding, DNA binding, and for transcriptional activation (16, 17). The DNA-binding domains, which contain about 70 residues, have eight conserved cysteine residues (Figure 12). Biochemical studies showed that a peptide from the DNA-binding do-

main of the glucocorticoid receptor could fold in the presence of zinc, and that this peptide specifically recognized the appropriate DNA-binding site (99).

Since the steroid receptors contain two sets of four cysteines, it had been proposed that this region would form a pair of "zinc fingers." However, the sequence patterns seen in the steroid receptors are very different than those found in the TFIIIA-like zinc fingers (cf. Figures 8 and 12), and it was clear that the steroid receptors formed a distinct structural motif (100). NMR studies of the DNA-binding domains from the glucocorticoid and estrogen receptors (101, 102) revealed that each of these peptides folds into a single globular domain with a pair of α-helices. The two extended helices are roughly perpendicular and are held together by hydrophobic contacts. A zinc ion binds near the start of each helix and holds a peptide loop against the N-terminal end of the helix.

Recently, a crystal structure for a complex of the glucocorticoid receptor has been reported (103). This structure shows that the peptide binds as a dimer, even though NMR studies had shown that the peptide exists as a monomer in solution (101). The first helix of each subunit fits into the major groove, and side chains from this helix make contacts with the edges of the base pairs. The second major helix provides phosphate contacts with the DNA backbone and provides the dimerization interface. The crystal structure also gives more information about the loop regions that precede the major helices. The loop of the N-terminal finger contains a short segment of antiparallel β structure, while the C-terminal finger contains a distorted α-helix.

Leucine Zipper and Helix-Loop-Helix

The leucine zipper (104) and helix-loop-helix proteins (105) have important roles in differentiation and development. They also are interesting because they illustrate the important roles that heterodimer formation can play in the regulation of gene expression.

The leucine zipper motif was first discovered as a conserved sequence pattern in several eukaryotic transcription factors (104), and it now is clear that this motif appears in a wide variety of transcription factors from fungi, plants, and animals (see references in Ref. 18). The DNA-binding domains of these leucine zipper proteins generally contain 60–80 residues (104, 106) and contain two distinct subdomains: the leucine zipper region mediates dimerization, while a basic region contacts the DNA.

Leucine zipper sequences are characterized by a heptad repeat of leucines over a region of 30–40 residues (104; see Figure 13). There also tends to be a conserved repeat of hydrophobic residues (often Val or Ile) occurring three residues to the N-terminal side of the leucines (104). Biochemical experiments suggested that the leucine zipper region forms two parallel α-helices in a coiled-coil arrangement (111), and a high-resolution structure of

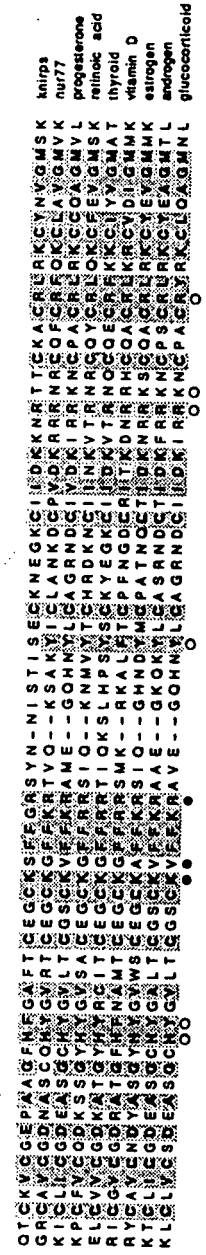

Figure 12 Alignment of steroid receptor sequences. Positions of the conserved cysteines (C) are shaded, as are positions that are generally acidic (D, E), basic (H, R, K), hydrophobic (A, C, V, I, L, M, F, Y), or small (G, A). The positions of secondary structure, backbone contacts (O), and major groove contacts (●) are from the structure of the glucocorticoid complex (103). The alignment is based upon that of Evans (16). For sequence references see Evans (16) and references therein, and (96–98).

Figure 13 Alignment of basic region-leucine zipper sequences. Positions that are generally conserved, basic (R, K), or hydrophobic (A, C, V, I, L, M, F, Y) are shaded. The position of the leucine zipper α-helix is from O'Shea et al (112). The alignment is based upon that of Kerppola & Curran (18). For sequence references, see (104, 106–110) and references therein. The sis-A sequence was determined by J. Erickson and T. Cline (unpublished).

this region from GCN4 has recently been reported (112). The coiled-coil arrangement readily explains the heptad repeat of the leucines and the offset heptad repeat of hydrophobic residues. These residues form the buried subunit interface of the coiled-coil dimer. Coiled-coils have a periodicity of 3.5 residues per helical turn, and thus every seventh residue is in the same structural environment.

The basic region (which contains about 30 residues) is rich in arginines and lysines, but also contains other residues that are conserved throughout the family or in particular subfamilies (18; Figure 13). Swapping basic regions and zipper regions from different proteins shows that the basic region is primarily responsible for the sequence preferences of the leucine zipper proteins (113). In fact, the basic region of GCN4 (without the leucine zipper!) can bind to its DNA site specifically as long as a disulfide bond is added to allow dimer formation (114). No crystal structure or NMR structure is available for the basic region, but it appears that this region is disordered in solution and α-helical in the complex. Peptides corresponding to this region are only about 25% helical in solution but are nearly 100% helical when bound to DNA (114–117). "Affinity cleavage" and protection experiments suggest that the basic regions are symmetrically positioned in the major groove (118, 120).

No direct structural data are available, but two closely related models for the complex have been proposed (115, 119). In these models, the peptide is shaped like a Y. The parallel leucine zipper region forms the stem of the Y, and the basic region forms α-helices that extend off like the arms of the Y. In each model, the N-terminal end of the two-fold symmetric leucine zipper fits over the center of the binding site, and the helices from the basic region extend in opposite directions along the major grooves of the DNA. In the "scissors grip" model, the α-helix from the basic region is kinked so that it can follow the curve of the major groove (119). In the "induced helical fork" model, this helix is straight and thus extends away from the DNA after contacting three or four base-pairs (115).

Leucine zipper proteins can form heterodimers, and these mixed dimers have important roles in regulating the biological activity of the bZIP proteins. For example, the active transcription factor AP-1 consists of one Fos protein and one Jun protein (121, 122). Heterodimer formation has several different roles in the leucine zipper family. Heterodimers can limit activity: CREB, a cAMP response regulator, is antagonized by formation of heterodimers with CREM (123). Heterodimers may also acquire new DNA-binding specificities (124) and thus be targeted to different sites than homodimers. In other cases, heterodimer formation may allow for different combinations of activation and/or repression domains and thus change the regulatory properties of a molecule bound at a fixed DNA site.

The helix-loop-helix (HLH) proteins (105, 125) appear to have some similarities with the leucine zipper family. Like the leucine zipper proteins, the HLH proteins have a basic region that contacts the DNA and a neighboring region that mediates dimer formation (126). Based upon sequence patterns, it has been proposed that this dimerization region forms an α-helix, a loop, and a second α-helix (105, Figure 14). The sequence of the basic region has some similarities with that of the basic region of the leucine zipper proteins (127), but it is not known whether these regions have similar three-dimensional structures.

Like the leucine zipper proteins, the HLH proteins have many important roles in differentiation and development, and their activity is modulated by heterodimer formation. Thus the MyoD protein, which appears to be the primary signal for differentiation of muscle cells, binds DNA most tightly when it forms a heterodimer with the ubiquitously expressed E2A protein (128). There also is a cellular protein that contains an HLH region—without the basic region—and apparently acts as a negative regulator by forming inactive heterodimers with MyoD (129). Heterodimer formation is used in many different ways in this family of proteins, mixing positive activators, ubiquitously expressed proteins, and negative regulators to modulate gene activity (130). There may be yet other families of proteins that use similar themes of homodimer and heterodimer formation to regulate activity. For example, studies of AP-2 led to proposals about a helix-span-helix class of proteins (131).

β-Sheet Motifs

Most of the major families of DNA-binding proteins that have been structurally characterized bind with an α-helix in the major groove. The MetJ, Arc, and Mnt repressors are interesting because they belong to a family of prokaryotic regulatory proteins that uses an antiparallel β-sheet for DNA binding (132, 133). The crystal structure of MetJ is known, and a preliminary description of the DNA complex has been reported (133, 134). The structure of Arc, both in solution (132) and in a crystal (U. Obeysekare, C. Kissinger, R. Sauer, and C. Pabo, unpublished), has also been determined. MetJ (which contains 104 residues) is significantly larger than Arc (which contains 53 residues), but both proteins form dimers in solution and the structures are homologous in a region that contains a β-strand and two α-helices. In the protein dimers, the β-strands pair to form an antiparallel β-sheet, while the α-helical regions pack against the sheet and against each other to stabilize the dimer.

MetJ binds as a tetramer to an 18-base-pair DNA site (134). In this complex, a dimer binds to each half-site, and each half-site contains a two-fold symmetry axis that is coincident with the two-fold axis of the β-sheet. The β-sheet fills the major groove, and side chains on the exposed

Figure 14 Alignment of basic region helix-loop-helix sequences. Positions that are generally conserved, acidic (D, E), basic (R, K), or hydrophobic (A, C, V, I, L, M, F, Y) are shaded. The sequences, alignment, and positions of potential secondary structure are from Murre et al (105) and references therein.

Figure 15 Alignment of β-sheet DNA-binding protein sequences. Positions that are generally conserved, acidic (D, E), basic (R, K), or hydrophobic (A, C, V, I, L, M, F, Y, W) are shaded. The secondary structure is that found in MetJ (134). The positions of major groove contacts (●) and backbone contacts (○) are from the MetJ complex (133). The alignment is from Breg et al (132). For sequence references, see (132) and (141) and references therein.

face of the sheet contact the base pairs (133). One lysine from each β-strand contacts a guanine, and a neighboring threonine on each strand contacts an adenine. Residues from the N-terminal end of α-helix 2 make backbone contacts. The MetJ tetramer is stabilized by interactions between α-helix 1 on one dimer and α-helix 1 on the other dimer.

It seems likely that Arc and Mnt interact with DNA in a way that is fundamentally similar to MetJ's interaction with DNA. Both proteins bind their operators as tetramers (136, 137), and genetic experiments have implicated the β-sheets in DNA recognition (135, 138, 139). However, the half-sites of the *arc* and *mnt* operators are not two-fold symmetric, and thus some breakdown of symmetry (in comparison with the fully symmetric arrangement in the MetJ complex) would need to occur. Biochemical experiments also show that Mnt uses an N-terminal arm to wrap around and contact the center of its operator (140). These interactions, which supplement the β-sheet contacts made by Mnt, seem generally similar to those made by the N-terminal arm of λ repressor.

Although Arc and Mnt show significant sequence similarities, their relationship with MetJ was only clearly established after structural studies of MetJ and Arc (132, 134). Even when the structures are used to align the sequences, there are relatively few positions at which a specific residue is conserved (Figure 15). Searches for additional members of this family, based upon hydrophobicity patterns, have identified the TraY proteins as probable relatives (141). However, given the limited sequence similarity among known members of this family, there could be other members that are simply too difficult to detect based upon sequence searches.

There appears to be at least one other family of regulatory proteins that uses β-sheets for DNA recognition. Although no structure is available, several arguments suggest that the *E. coli* IHF protein uses β-sheets for site-specific recognition (142). IHF is homologous with the bacterial HU protein, and crystallographic analysis of HU shows that it contains a pair of antiparallel β-arms that might participate in DNA recognition (143, 144). Based on chemical protection, sequence homology, and other biochemical data, a model has been proposed with β-sheet regions from IHF fitting in the minor groove of a sharply bent DNA site (142).

Other Families

There are a number of additional families of DNA-binding proteins that have been identified in studies of transcription factors, and presumably more families will be discovered in the future. As mentioned above, there are several other families (distinct from the TFIIIA-like fingers and the steroid receptors) that use zinc or other metal ions to stabilize their structures. For example, the cysteine-rich motif in the yeast GAL4 repressor has a distinct

sequence pattern (Cys-X2-Cys-X6-Cys-X6-Cys-X2-Cys-X6-Cys) and is representative of a separate family of "fungal fingers" (145). NMR studies have shown that GAL4 contains a distinctive binuclear metal cluster (146), and this set of proteins has recently been referred to as the "zinc cluster" family (15). The "CCHC" motif, which has a sequence pattern of the form Cys-X2-Cys-X4-His-X4-Cys, is yet another metal-binding motif involved in nucleic acid recognition (147). This motif occurs in a set of retroviral gag proteins, and the structure, as determined by 2D NMR for a peptide from the HIV gag protein, is similar to the iron-binding region of rubredoxin (148).

There are several other families of DNA-binding proteins with cysteine-rich motifs that may serve as metal-binding domains. For example, a cysteine-rich motif (with the general form Cys-X-Asn-Cys-X17-Cys-Asn-X-Cys) was discovered in the GATA factor, which is specifically expressed in erythroid cells (149, 150). RAG-1 and RAD18 are prototypical members of a distinctive family of DNA-binding proteins that contain another cysteine-rich motif (151). The LIM motif is a distinctive cysteine-rich region that occurs on the N-terminal side of the homeodomain in several regulatory proteins from *Caenorhabditis elegans* (152). The ACE1 transcription factor of yeast (which may be related to metallothionein) has a cysteine-rich region that appears to form a cluster with 6–7 copper ions (153).

Although cysteine-rich and histidine-rich motifs are easily detected in sequence comparisons (these residues occur relatively rarely in most proteins) and have received considerable attention, several other families of DNA-binding proteins have been characterized recently. One prominent family (that was mentioned previously) is the set of POU proteins, which contain a homeodomain and a POU-specific domain that also binds DNA (73–75). There also are families of transcriptional regulatory proteins related to c-ets (154), to the nonhistone high mobility group proteins (155 and references therein), to the serum response factor (156 and references therein), to c-myb (157 and references therein), to NF-κB, and to the *rel* oncogene (158–160). Sequence comparisons also suggest that there are groups of DNA-binding proteins related to the paired domain from *Drosophila* (161, 162), to the homeotic gene fork head (163, 164), and to a set designated as the TEA domain (165). At present, no detailed structural information is available about these families. Obviously, a thorough analysis will require information on the three-dimensional structure, biological distribution, and regulatory roles of each of these families.

PRINCIPLES OF RECOGNITION

Helices in Recognition

Now that a set of protein-DNA complexes has been solved, we can take a broader look at the problem of protein-DNA recognition and search for useful

generalizations. Most of the well-characterized families of DNA-binding proteins use α-helices to make base contacts in the major groove. Although β-sheets or regions of extended polypeptide chain can also make contacts, α-helices are used much more frequently. In fact, the use of α-helices in site-specific recognition is so widespread that there is some danger of overemphasizing their role or of misunderstanding the different ways in which an α-helix can be used. As mentioned above, referring to a particular helix as a "recognition helix" can easily be misleading. There is no evidence that an isolated helix from any of the known motifs can bind DNA in a site-specific fashion, and no regulatory system has been discovered that uses an isolated helical peptide (i.e. a single α-helix) as a site-specific DNA-binding protein. In every reported complex involving α-helical motifs, it appears that the overall binding specificity results from a set of interactions, with some contacts from a "recognition" helix and some from surrounding regions of the protein.

Although the orientation of α-helices with respect to the major groove may be conserved within a given family or subfamily of DNA-binding proteins, there is no unique way of placing an α-helix in the major groove. Thus, the orientations of the α-helices are significantly different when complexes from different DNA-binding families are compared (e.g. λ repressor, the engrailed homeodomain, and the zif268 zinc fingers). Even within a given family, the orientations of the α-helices can be significantly different (as with λ repressor and Trp repressor). The main point is that the overall shape and dimensions of an α-helix allow it to fit into the major groove in a number of related, but significantly different ways. Some helices lie in the middle of the major groove and have the axis of the α-helix approximately tangent to the local direction of the major groove (i.e. at an angle of 32° with respect to the plane that is perpendicular to the helical axis of the DNA). Other α-helices are tipped at different angles, varying at least 15° in each direction, and some are arranged so that only the N-terminal portion of the α-helix fits completely into the major groove. It appears that surrounding regions of the proteins (and not just the sequences of the "recognition helices") help to determine how these α-helices are positioned in the major groove.

Interactions with Bases

Contacts with the bases play a crucial role in site-specific binding, and the known complexes contain a rather diverse set of base contacts. Structural studies have revealed (*a*) direct hydrogen bonds between the protein side chains and the bases, (*b*) occasional hydrogen bonds between the polypeptide backbone and the bases, (*c*) hydrogen bonds mediated by water molecules, and (*d*) hydrophobic contacts. Although a few base contacts occur in the minor groove (such as those made by the N-terminal arm of the homeodomain), the major groove appears to be more important for site-specific

recognition. This is not surprising, since the major groove is larger and more accessible in B-DNA and has more potential sites for hydrogen bonding and hydrophobic contacts. Analyzing the pattern of hydrogen-bonding sites along the edge of the base pairs also suggested that major groove contacts would provide a more reliable basis for sequence-specific recognition (166). In the minor groove, the O2 of thymine and the N3 of adenine occupy very similar positions, and may be difficult to distinguish since both are hydrogen bond acceptors.

A comparison of the base contacts that have been observed in various complexes shows that there is no simple "recognition code"; i.e. there is no one-to-one correspondence between the amino acid side chains and the bases they contact. Figure 16 shows some of the hydrogen-bonding interactions that have been observed between side chains and bases in the known complexes. Some side chains are used to contact more than one kind of base, and some bases are contacted by a variety of side chains. No simple pattern or rule describes all of these contacts, although the table suggests that hydrogen bonds with purines (particularly with guanine!) may have an especially important role in recognition. What accounts for this diversity of contacts? Even when the same secondary structure, such as an α-helix, is used for base recognition, the precise position and orientation of the α-helix may determine what set of side chain–base interactions are possible for a particular residue. Further levels of complexity are added when we consider side chain–side chain interactions, residues that contact more than one base pair, and contacts that involve bridging water molecules or ions. There also are cases—as observed with λ repressor's N-terminal arm—in which the polypeptide backbone of the protein hydrogen bonds with the edge of the base pairs (50). Overall, the variety and structural complexity of the contacts involved in protein-DNA interactions rival those of the contacts involved in protein structure and folding.

In spite of these complexities, some organizing principles emerge when we compare closely related structures, such as the λ and 434 repressors or the engrailed and α2 homeodomains. In these cases, we find that conserved residues at the protein-DNA interface make conserved contacts with the DNA. This was first noted when comparing the λ and 434 repressors, which have three conserved residues on the DNA-binding face of the HTH unit (49). Each of these conserved residues makes conserved contacts with the DNA. A glutamine at the start of the HTH unit makes conserved contacts with the DNA backbone, a glutamine at the start of the "recognition helix" makes conserved hydrogen bonds with an adenine, and an asparagine at the end of the HTH unit makes a conserved contact with the DNA backbone. Comparing the engrailed and α2 homeodomain–DNA complexes shows that conserved residues at the protein-DNA interface also make conserved contacts with the

	A	T	C	G
Arg		CAP		GR Trp zif CAP
Asn	en			lambda
Gln	lambda 434	434		434
Glu			CAP	
His				zif
Lys				GR lambda MetJ
Ser				lambda
Thr	MetJ			

Figure 16 Chart summarizing hydrogen bonds between side chains and bases in the major groove for representative protein-DNA complexes. Contacts are summarized for the *E. coli* CAP protein (56), the *Drosophila* engrailed homeodomain [en, (66)], the glucocorticoid receptor [GR, (103)], the 434 repressor (48), the λ repressor (47), the *E. coli* MetJ repressor (133), the *E. coli* Trp repressor (54), and the zif268 zinc fingers (83).

DNA (66, 67). The amino acid sequences of these two homeodomains are only 27% identical, but six conserved residues—Phe8, Tyr25, Trp48, Arg53, Lys55, and Lys57—make conserved contacts with the DNA backbone. Asn51, which is invariant in the homeodomain family, makes a conserved set of hydrogen bonds with an adenine. Comparing fingers 1, 2, and 3 of the zif268 complex (83) again shows that conserved residues at the protein-DNA interface make conserved contacts with the DNA. There are several conserved contacts with the DNA backbone and several conserved contacts with the bases.

These comparisons suggest that a given family or subfamily of DNA-binding proteins may have—in addition to the conserved folding motif that characterizes the family—a conserved "docking mechanism" and a conserved set of contacts. This may provide a useful simplifying principle when thinking about recognition. Contacts may be more predictable if we focus attention on a particular position within a particular structural motif. In some sense, the position and orientation of the polypeptide backbone with respect to the DNA determine the "meaning" of a particular side chain (167). The fact that a side chain can be presented in different ways by different structural motifs (or even by different positions within a given motif) may be one of the reasons that there is no simple "recognition code" (Figure 16). Since structure is central to recognition, the folding of the polypeptide and the overall docking arrangement help to determine which base contacts are possible in any given situation. In this context, it is intriguing that the conserved base contacts that have been observed all involve the types of side chain–base interactions that had been predicted by Seeman, Rosenberg, and Rich (166). These include the interaction of arginine with guanine, and the interaction of glutamine or asparagine with adenine. Since each of these side chain–base contacts involves a pair of hydrogen bonds, they may play an especially important role in site-specific recognition.

Contacts with the DNA Backbone

Even a cursory examination of protein-DNA complexes suggests that contacts with the DNA backbone play an integral role in site-specific recognition. In the known complexes, roughly half of all the hydrogen bonds involve contacts with the DNA backbone. Almost all of these contacts involve the phosphodiester oxygens. A few hydrophobic interactions with sugar rings in the DNA backbone have been reported, but these are much less common and presumably far less important for recognition.

As observed with side chain–base interactions, there does not appear to be any simple "rule" or pattern describing which residues are used for backbone contacts. Examining the published complexes shows that many different side chains, and even the -NH of the polypeptide backbone, can hydrogen bond to

the phosphodiester oxygens. Given the variety of contacts that have been observed, it seems likely that any basic or neutral hydrogen-bonding side chain can be used to contact the phosphodiester oxygens. We assume that aspartic acid and glutamic acid are used less frequently because of the unfavorable electrostatic interactions, but even these side chains could contact the phosphates via bridging divalent cations such as Mg^{2+} or Ca^{2+}. Although lysine and arginine are used in a number of cases, their favorable electrostatic interactions with the phosphates may be partially offset by the inherent flexibility of their long side chains. In some complexes (e.g. λ repressor), short polar side chains and the peptide -NH seem to play a more important role in contacting the backbone. These contacts may provide more stereospecificity than those mediated by arginines or lysines.

The sheer number of hydrogen bonds with the phosphodiester oxygens, and the exquisite hydrogen-bonding networks that have been observed (47, 48, 54), suggest that these contacts are very important for site-specific recognition. We still do not understand the exact role of backbone contacts, but two major roles appear possible: (*a*) Backbone contacts may serve as "fiducial marks" that help hold the protein against the bases in a fixed arrangement and thereby enhance the specificity of the side chain–base interactions. Without these backbone contacts to orient the protein in the major groove and help establish a fixed register for the interactions, the "recognition helix" might slip or shift in a way that would allow it to hydrogen bond with inappropriate sites. (*b*) Sequence-dependent variations in the DNA structure may also be important (as discussed in the next section). To the extent that base sequence determines the structure of the DNA—and therefore the most favorable positions for the phosphodiester oxygens—contacts with the DNA backbone may allow indirect recognition of the sequence.

Role of DNA Structure in Recognition

Any thorough discussion of binding specificity requires consideration of variations in DNA structure and/or flexibility. There are several ways in which the local structure of the DNA could influence specificity: (*a*) There could be sequence-dependent influences on the structure of a given binding site, causing it to have an average structure that deviates from canonical B-form DNA (54, 168). Specificity would be enhanced if a protein were designed to recognize a stable local structure such as a DNA bend or kink. (*b*) There could be sequence-dependent effects on the flexibility of the DNA, which could contribute to specificity if the protein distorts the DNA (possibly by bending) as it binds. Even if the average solution structure of the DNA were linear, bending could contribute to recognition if the correct site were more flexible (i.e. more easily bent) than canonical B-form DNA.

There now are a number of examples in which repressors bind to bent sites

on the DNA, but it still is difficult to evaluate the contributions of these effects to site-specific recognition. Thus significant bending of the DNA occurs in complexes with the Trp repressor (54), 434 repressor (48), 434 Cro protein (53), λ repressor (47), and λ Cro protein (55). In each case, the DNA bends as if it were beginning to "wrap around" the HTH units of the proteins. An even more dramatic bend of 90° occurs in the *E. coli* CAP complex (56). It seems plausible that these bends (especially the dramatic bend observed in the CAP complex) make some contribution to specificity, but it still is difficult to evaluate the magnitude of these effects. The main problems stem from the difficulty of determining the average structure of the free DNA and the energetic cost of bending the DNA. It would be useful to have structures of the free DNA-binding sites, but solution structures determined by 2D NMR do not provide sufficient structural detail (169), and crystal structures of the free DNA may be seriously perturbed by crystal packing forces (170). Energetic analyses present an even more serious problem. To understand the energetic contributions to recognition, we really would need to compare the energetic costs of distorting the correct site (which could be zero if this site is bent in solution) with the energetic costs of distorting other sites. At present, there does not appear to be any reliable way of determining these energies.

Trying to understand the role of contacts with the DNA backbone raises several related problems about recognition. One involves the structural relationship between specific and nonspecific binding. In particular, do DNA-binding proteins make the same backbone contacts when they bind to nonspecific DNA? This is crucial to understanding the role of backbone contacts in recognition, but unfortunately there is very little structural information about nonspecific complexes. [A symmetric DNA site, with an altered spacing between the half-sites, was used when cocrystallizing the DNA-binding domain of the glucocorticoid receptor, and one-half of this complex provides us with our first glimpse of a nonspecific complex (103).] Since biological specificity only results from the free energy differences between specific and nonspecific binding, we can never really understand site-specific binding without understanding the nonspecific "reference state."

Considering the role of contacts with the phosphodiester oxygens raises broader problems involved in any attempts to assign separate energetic contributions to particular contacts or groups within protein-DNA complexes. One can easily study the binding of a mutant protein or of a DNA molecule in which a particular functional group has been removed, but it is not correct to assume that the $\Delta\Delta G$ measured for such a variant only reflects the contribution of this particular contact. First, one would need to know that the mutation had not altered the structure of the complex. Changing even one contact could allow subtle changes throughout the complex. Second, it seems likely that individual contacts—including base contacts and backbone contacts—are

intimately coupled. One would need to study many different variants and test various sets of changes in any attempt to assess properly the intrinsic and cooperative energies of the system. Careful studies of the structures and energies of mutant complexes are sure to yield useful insights, but the reductionist approach can only be taken so far. Ultimately, recognition involves a set of contacts with a set of sites on the DNA, and attempting to describe the energy of the complex as a sum of separate, discrete contributions may be meaningless.

General Principles of Site-Specific Recognition

Although the diversity of known DNA-binding motifs and contacts suggests that there are no simple rules or patterns describing site-specific recognition, comparing the known complexes allows us to make a few broad generalizations.

1. Site-specific recognition always involves a set of contacts with the bases and with the DNA backbone.
2. Hydrogen bonding is critical for recognition (although hydrophobic interactions also occur). A complex typically has on the order of 1–2 dozen hydrogen bonds at the protein/DNA interface.
3. Side chains are critical for site-specific recognition. There are instances in which the peptide backbone makes hydrogen bonds with bases or the DNA backbone, but side chains make most of the critical contacts.
4. There is no simple one-to-one correspondence between side chains and the bases they contact. It appears that the folding and docking of the entire protein help to control the "meaning" that any particular side chain has in site-specific recognition.
5. Most of the base contacts are in the major groove. Contacts with purines (which are larger and offer more hydrogen-bonding sites in the major groove) seem to be especially important.
6. Most of the major motifs contain an α-helical region that fits into the major groove of B-form DNA. There are examples of β-sheets and/or extended regions of polypeptide chain that play critical roles in certain proteins, but base contacts from these regions appear to be less common.
7. Contacts with the DNA backbone usually involve hydrogen bonds and/or salt bridges with the phosphodiester oxygens.
8. Multiple DNA-binding domains usually are required for site-specific recognition. The same motif may be used more than once, as occurs when the active binding species is a homodimer or heterodimer, or when a single polypeptide contains tandem recognition motifs. Different motifs (an extended arm and a HTH unit; a homeodomain and POU-specific domain, etc) may also be used in the same complex.

9. Recognition is a detailed structural process. Hydration can play a critical role in recognition; sequence-dependent aspects of the DNA structure may also be important.

DIRECTIONS FOR FUTURE RESEARCH

The past few years have brought exciting progress in the study of protein-DNA recognition, but important questions remain. Although many of the problems are interrelated, some of the most critical questions involve issues of structure, energy, evolution, gene regulation, and protein design.

Structure Recent structural studies have given us a much better understanding of protein-DNA interactions, but much work remains. Obviously, it will be important to solve the structures of complexes that contain some of the other major DNA-binding motifs (leucine zippers, helix-loop-helix proteins, etc). It also will be important to determine the structures of some intact regulatory proteins. We need to know the structures of other domains to understand transcriptional activation and other activities of the intact proteins. It also is possible that neighboring domains affect the specificity, affinity, or accessibility of the DNA-binding domain. Finally, there is a continuing need to examine more examples of the "known" DNA-binding motifs. The fact that the prokaryotic HTH unit and the eukaryotic homeodomain bind in such different ways should serve as something of a warning. Other surprises may await us, and it certainly is possible that there are multiple ways to use a zinc finger or a homeodomain in DNA recognition.

Modeling presents additional challenges for future structural work. Obviously, it would be a tremendous advance if we could reliably model protein-DNA interactions. Does the wealth of new structural data allow us to improve our methods for modeling? Is it now possible to model reliably homologous complexes? Will the development of improved potential functions and/or the development of new computational strategies eventually allow us to model other protein-DNA complexes?

Studies of DNA-binding proteins certainly will benefit from any improvements that can be made in the methods for correlating amino acid sequence and protein structure. Bowie & Eisenberg (171) have recently presented a method that uses patterns derived from protein crystal structures to search for sequences that could reasonably be expected to fold in the same manner. It will be interesting to see if this method can identify new members for any of the major DNA-binding families for which structures are available.

Energy A real chemical and physical understanding of recognition also will require a better understanding of the energetics of protein-DNA interactions.

Analyzing the energies either experimentally or theoretically is complicated because recognition always involves a set of contacts. It is difficult to "dissect" these interactions in a way that assigns specific energetic contributions to individual contacts. These problems are further complicated by the fact that we usually are interested in the energetic differences between contacts that occur in a specific complex and contacts that occur with nonspecific DNA. (These difference in the binding energy account for the biological specificity.) Modern computational methods, such as those involving free energy perturbation (172), may be useful, but these studies are difficult and computationally expensive. These calculations also require very high-resolution crystal structures as a starting point, and we need more detailed structural information about nonspecific complexes before we have any chance of understanding the energetic differences that are responsible for site-specific recognition. Obviously, these energetic problems are complicated. At some point, they overlap with more general and fundamental questions about protein folding and macromolecular recognition.

Other important questions—such as trying to understand why so many proteins use α-helices for recognition—involve a combination of structural and energetic issues. Is there some inherent advantage to using an α-helix in recognition? Clearly, the size and shape of an α-helix make it complementary to the major groove, but a β-sheet can also fit in the major groove. Why do relatively few proteins use β-sheets for site-specific binding? Is the α-helix more "rigid" in some way that helps to enhance the specificity of the contacts? Are there more distinct ways of arranging an α-helix in the major groove and thus more sequences that helical proteins can recognize?

Trying to understand the role of flexible regions in site-specific recognition also involves structural and energetic issues. Why are flexible or disordered regions (like λ repressor's N-terminal arm and the basic region of the leucine zipper proteins) sometimes used for DNA-binding? Do these regions of the polypeptide need to be flexible to allow rapid binding? Would a rigid protein bind too slowly and be released too slowly? Alternatively, do these disorder $\rightarrow$ order transitions provide an entropic way of limiting the overall binding energy while retaining a set of protein-DNA contacts that are essential for specificity?

Evolution Although the known families provide a plausible way of grouping DNA-binding proteins, there still are many interesting questions about subfamily and superfamily relationships. We cannot yet construct an accurate genealogy for the families of DNA-binding proteins or understand the structural significance of all of the conserved sequence patterns. For instance, are the prokaryotic HTH motif and the eukaryotic homeodomain related by divergent or convergent evolution? Why are these structures so similar and yet

used in different ways when binding to DNA? What is the significance of the various homeodomain subfamilies that have been noted? Are the various families of zinc finger proteins related to each other in any meaningful way? Are the leucine zipper and helix-loop-helix proteins structurally related? As we look to the more distant past and try to visualize the earliest stages in the evolution of modern gene regulation, we also wonder: What are the smallest structural units that could plausibly have given some selective advantage during evolution? Could an isolated HTH unit or a single zinc finger have any meaningful regulatory role?

Gene Regulation Studies of protein-DNA recognition are closely linked to problems of gene regulation, but additional work is needed to really understand the connections between recognition and regulation. A careful analysis of the binding energies is important, since we would like to understand the physical basis for the differential regulation of gene expression. How do homeodomains distinguish among closely related binding sites? What role do accessory proteins play in site-specific binding? How much energy is contributed by protein-protein interactions? How does competition with other regulatory proteins and competition with nucleosomes affect binding? Clearly, additional structural work will be needed to help us understand the structural basis for positive and negative control. As mentioned above, it becomes extremely important to have structural information about the intact proteins so that we can see the domains responsible for positive and negative regulation.

Another question involves the apparent association between certain structural motifs and specific biological roles. Are certain motifs, such as the helix-loop-helix motif, particularly well adapted for regulating differentiation and development? Could other families of proteins, such as zinc fingers, have been used as effectively in these roles? It also seems possible that certain motifs are particularly well suited for recognizing certain base sequences. Does each motif have characteristic DNA sequences that are most suitable for the binding site? Are some motifs more "adaptable" than others and thus suited for recognizing a wider range of binding sites?

Design Attempts to design DNA-binding proteins provide another exciting challenge for future research. This will provide a rigorous test of our understanding, and could have tremendous practical benefits since new proteins might be used for research, diagnosis, or even gene therapy. When working on design, it may make sense to use known motifs from the major families of DNA-binding proteins. These provide an attractive starting point because each of these motifs is based on a stable secondary structure and has a good way of packing against the DNA. Moreover, the existence of large families of DNA-binding proteins suggests that these motifs offer highly successful and

adaptable frameworks for protein-DNA recognition. Pavletich & Pabo (83) have suggested that the zinc finger motif may provide a particularly attractive framework for the design of new DNA-binding proteins. Since the zinc fingers recognize an asymmetrical DNA sequence, there is no need to restrict the design process to work with symmetric target sites. It also is possible that the modular nature of the zinc finger–binding units will allow simplification of the design problem. Since each finger makes its primary contacts in a three-base-pair region, it might be possible to design or select fingers that recognize each of the 64 possible base triplets. If one can "mix and match" these fingers, it should be possible to design proteins that recognize any desired base sequence.

Ultimately, the problems of structure, energy, evolution, gene regulation, and design are closely related. We have made significant progress, but much work is needed before we have a real understanding of protein-DNA interactions. We should not be content with a superficial understanding that merely allows us to rationalize a few key observations. We need a deeper understanding with real explanatory power—an understanding that will let us predict how other proteins bind and that will allow us to design new DNA-binding proteins. Exciting challenges lie ahead.

ACKNOWLEDGMENTS

We thank our colleagues for helpful discussions and for permission to cite unpublished information, and thank Kristine Kelly for assistance in preparation of this manuscript. Work in our laboratories was supported NIH grant AI-16892 (R.T.S.), by NIH grant GM-31471 (C.O.P.), and by the Howard Hughes Medical Institute (C.O.P.).

Literature Cited

1. Steitz, T. A. 1990. *Q. Rev. Biophys.* 23:205–80
2. Johnson, P. F., McKnight, S. L. 1989. *Annu. Rev. Biochem.* 58:799–839
3. Mitchell, P. J., Tjian, R. 1989. *Science* 245:371–78
4. Struhl, K. 1989. *Trends Biol. Sci.* 14:137–40
5. Frankel, A. D., Kim, P. S. 1991. *Cell* 65:717–19
6. He, X., Rosenfeld, M. G. 1991. *Neuron* 7:183–96
7. Harrison, S. C. 1991. *Nature.* 355:715–19
8. Brennan, R. G., Matthews, B. W. 1989. *J. Biol. Chem.* 264:1903–6
9. Harrison, S. C., Aggarwal, A. K. 1990. *Annu. Rev. Biochem.* 59:933–69
10. Brennan, R. G. 1991. *Curr. Opin. Struc. Biol.* 1:80–88
11. Berg, J. M. 1986. *Science* 232:485–87
12. Klug, A., Rhodes, D. 1987. *Trends Biochem. Sci.* 12:464–69
13. Kaptein, R. 1991. *Curr. Opin. Struc. Biol.* 1:63–70
14. Berg, J. M. 1990. *Annu. Rev. Biophys. Biophys. Chem.* 19:405–21
15. Vallee, B. L., Coleman, J. E., Auld, D. S. 1991. *Proc. Natl. Acad. Sci. USA* 88:999–1003
16. Evans, R. M. 1988. *Science* 240:889–95
17. Beato, M. 1989. *Cell* 56:335–44
18. Kerppola, T. K., Curran, T. 1991. *Curr. Opin. Struc. Biol.* 1:71–79
19. Scott, M. P., Tamkun, J. W., Hartzell, G. W. 1989. *Biochim. Biophys. Acta* 989:25–48
20. Gehring, W. J., Muller, M., Affolter, M., Percival-Smith, A., Billeter, M., et al. 1990. *Trends Genet.* 6:323–29

21. Wright, C. V., Cho, K. W., Oliver, G., De Robertis, E. M. 1989. *Trends Biochem. Sci.* 14:52–56
22. Hayashi, S., Scott, M. P. 1990. *Cell* 63:883–94
23. Phillips, S. E. V. 1991. *Curr. Opin. Struc. Biol.* 1:89–98
24. Luisi, B. F., Sigler, P. B. 1990. *Biochim. Biophys. Acta* 1048:113–26
25. Sauer, R. T., Jordan, S. R., Pabo, C. O. 1990. *Adv. Prot. Chem.* 40:1–61
26. Pabo, C. O., Sauer, R. T. 1984. *Annu. Rev. Biochem.* 53:293–321
27. Arthur, A. K., Hoss, A., Fanning, E. 1988. *J. Virol.* 62:1999–2006
28. Kern, S. E., Kinzler, K. W., Bruskin, A., Jarosz, D., Friedman, P., Prives, C., et al. 1991. *Science* 252:1708–11
29. Anderson, W. F., Ohlendorf, D. H., Takeda, Y., Matthews, B. W. 1981. *Nature* 290:754–58
30. Ohlendorf, D. H., Anderson, W. F., Fisher, R. G., Takeda, Y., Matthews, B. W. 1982. *Nature* 298:718–23
31. McKay, D. B., Steitz, T. A. 1981. *Nature* 290:744–49
32. Pabo, C. O., Lewis, M. 1982. *Nature* 298:443–47
33. Steitz, T. A., Ohlendorf, D. H., McKay, D. B., Anderson, W. F., Matthews, B. W. 1982. *Proc. Natl. Acad. Sci. USA* 79:3097–100
34. Ohlendorf, D. H., Anderson, W. F., Lewis, M., Pabo, C. O., Matthews, B. W. 1983. *J. Mol. Biol.* 169:757–69
35. Kaptein, R., Zuiderweig, E. R. P., Sheek, R. M., Boelens, R., van Gunsteren, W. F. 1985. *J. Mol. Biol.* 182:179–82
36. Anderson, J. E., Ptashne, M., Harrison, S. C. 1987. *Nature* 326:846–52
37. Mondragon, A., Subbiah, S., Almo, S. C., Drottar, M., Harrison, S. C. 1989. *J. Mol. Biol.* 205:189–200
38. Wolberger, C., Dong, Y. C., Ptashne, M., Harrison, S. C. 1988. *Nature* 335:789–95
39. Mondragon, A., Wolberger, C., Harrison, S. C. 1989. *J. Mol. Biol.* 205:179–88
40. Schevitz, R. W., Otwinowski, Z., Joachimiak, A., Lawson, C. L., Sigler, P. B. 1985. *Nature* 317:782–86
41. Kostrewa, D., Granzin, J., Koch, C., Choe, H. W., Raghunathan, S., et al. 1990. *Nature* 349:178–80
42. Lamerichs, R. M., Padilla, A., Boelens, R., Kaptein, R., Ottleben, G., et al. 1990. *Proc. Natl. Acad. Sci. USA* 86:6863–67
43. Matthews, B. W., Ohlendorf, D. H., Anderson, W. F., Takeda, Y. 1982. *Proc. Natl. Acad. Sci. USA* 79:1428–32
44. Sauer, R. T., Yocum, R. R., Doolittle, R. F., Lewis, M., Pabo, C. O. 1982. *Nature* 298:447–51
45. Weber, I. T., McKay, D. B., Steitz, T. A. 1982. *Nucleic Acids Res.* 10:5085–102
46. Koch, C., Vandekerckhove, J., Kahmann, R. 1988. *Proc. Natl. Acad. Sci. USA* 85:4237–41
47. Jordan, S. R., Pabo, C. O. 1988. *Science* 242:893–99
48. Aggarwal, A. K., Rodgers, D. W., Drottar, M., Ptashne, M., Harrison, S. C. 1988. *Science* 242:899–907
49. Pabo, C. O., Aggarwal, A. K., Jordan, S. R., Beamer, L. J., Obeysekare, U. R., et al. 1990. *Science* 247:1210–13
50. Clarke, N. D., Beamer, L. J., Goldberg, H. R., Berkower, C., Pabo, C. O. 1991. *Science* 254:267–70
51. Pabo, C. O., Krovatin, W., Jeffrey, A., Sauer, R. T. 1982. *Nature* 298:441–43
52. Weiss, M. A., Sauer, R. T., Patel, D. J., Karplus, M. 1984. *Biochemistry* 23: 5090–95
53. Mondragon, A., Harrison, S. C. 1991. *J. Mol. Biol.* 219:321–34
54. Otwinowski, Z., Schevitz, R. W., Zhang, R-G., Lawson, C. L., Joachimiak, A., et al. 1988. *Nature* 335:321–29
55. Brennan, R. G., Roderick, S. L., Takeda, Y., Matthews, B. W. 1990. *Proc. Natl. Acad. Sci. USA* 87:8165–69
56. Schulz, S. C., Shields, G. C., Steitz, T. A. 1991. *Science* 253:1001–7
57. Koudelka, G. B., Harrison, S. C., Ptashne, M. 1987. *Nature* 326:886–88
58. Koudelka, G. B., Harbury, P., Harrison, S. C., Ptashne, M. 1988. *Proc. Natl. Acad. Sci. USA* 85:4633–37
59. Laughon, A., Scott, M. P. 1984. *Nature* 320:25–31
60. Shepherd, J. C. W., McGinnis, W., Carrasco, A. E., De Robertis, E. M., Gehring, W. J. 1984. *Nature* 310:70–71
61. Sauer, R. T., Smith, D. L., Johnson, A. D. 1988. *Genes Dev.* 2:807–16
62. Affolter, M., Percival-Smith, A., Muller, M., Leupin, W., Gehring, W. J. 1990. *Proc. Natl. Acad. Sci. USA* 87: 4093–97
63. Qian, Y. Q., Billeter, M., Otting, G., Muller, M., Gehring, W. J., et al. 1989. *Cell* 59:573–80
64. Billeter, M., Qian, Y., Otting, G., Muller, M., Gehring, W. J., et al. 1990. *J. Mol. Biol.* 214:183–97
64a. Phillips, C., Vershon, A., Johnson, A., Dahlquist, F. 1991. *Genes Dev.* 5:764–72
65. Otting, G., Qian, Y. Q., Billeter, M., Muller, M., Affolter, M., et al. 1990. *EMBO J.* 9:3085–92
66. Kissinger, C. R., Liu, B. S., Martin-Blanco, E., Kornberg, T. B., Pabo, C. O. 1990. *Cell* 63:579–90
67. Wolberger, C., Vershon, A. K., Liu, B., Johnson, A. D., Pabo, C. O. 1991. *Cell* 67:517–28
68. Garcia-Blanco, M. A., Clerc, R. G., Sharp, P. A. 1989. *Genes Dev.* 3:739–45
69. Hanes, S. D., Brent, R. 1989. *Cell* 57:1275–83
70. Treisman, J., Gonczy, P., Vashishtha, M., Harris, E., Desplan, C. 1989. *Cell* 59:553–62
71. Hanes, S. D., Brent, R. 1991. *Science* 251:426–30
72. Affolter, M., Percival-Smith, A., Muller, M., Billeter, M., Qian, Y. Q., et al. 1991. *Cell* 64:879–80
73. Rosenfeld, M. G. 1991. *Genes Dev.* 5:897–907
74. Ruvkun, G., Finney, M. 1991. *Cell* 64:475–78
75. Herr, W., Sturm, R. A., Clerc, R. G., Corcoran, L. M., Baltimore, D., et al. 1988. *Genes Dev.* 2:1513–16
76. Goutte, C., Johnson, A. D. 1988. *Cell* 52:875–82
77. Keleher, C. A., Passmore, S., Johnson, A. D. 1989. *Mol. Cell. Biol.* 9:5228–30
78. Kristie, T., Sharp, P. 1990. *Genes Dev.* 4:2383–96
79. Miller, J., McLachlan, A. D., Klug, A. 1985. *EMBO J.* 4:1609–14
80. Brown, R. S., Sander, C., Argos, P. 1985. *FEBS Lett.* 186:271–74
81. Ginsberg, A. M., King, B. O., Roeder, R. G. 1984. *Cell* 39:479–89
82. Krizek, B. A., Amann, B. T., Kilfoil, V. J., Merkle, D. L., Berg, J. M. 1991. *J. Am. Chem. Soc.* 113:4518–23
83. Pavletich, N. P., Pabo, C. O. 1991. *Science* 252:809–17
84. Berg, J. M. 1988. *Proc. Natl. Acad. Sci. USA* 85:99–102
85. Gibson, T. J., Postma, J. P. M., Brown, R. S., Argos, P. 1988. *Prot. Eng.* 2:209–18
86. Parraga, G., Horvath, S. J., Eisen, A., Taylor, W. E., Hood, L., et al. 1988. *Science* 241:1489–92
87. Lee, M. S., Gippert, G. P., Soman, K. V., Case, D. A., Wright, P. E. 1989. *Science* 245:635–37
88. Christy, B. A., Lau, L. F., Nathans, D. 1988. *Proc. Natl. Acad. Sci. USA* 85:7857–61
89. Nardelli, J., Gibson, T., Vesque, C., Charnay, P. 1991. *Nature* 349:175–78
90. Kadonaga, J. T., Carner, K. R., Masiarz, F. R., Tjian, R. 1987. *Cell* 51:1079–90
91. Mardon, G., Page, D. C. 1989. *Cell* 56:765–70
92. Kochoyan, M., Havel, T., Nguyen, D., Dahl, C., Keutmann, H., et al. 1991. *Biochemistry* 30:3371–86
93. Baldwin, A. Jr., LeClair, K., Singh, H., Sharp, P. 1990. *Mol. Cell Biol.* 10:1406–14
94. Reuter, G., Giarre, M., Farah, J., Gausz, J., Spierer, A., Spierer, P. 1990. *Nature* 344:219–23
95. Fasano, L., Roder, L., Core, N., Alexandre, E., Vola, C., et al. 1991. *Cell* 64:63–79
96. Nauber, U., Pankratz, M. J., Kienlin, A., Seifert, E., Klemm, U., Jaekle, H. 1988. *Nature* 336:489–92
97. Hazel, T. G., Nathans, D., Lau, L. F. 1988. *Proc. Natl. Acad. Sci. USA* 85:8444–48
98. Tilley, W. D., Marcelli, M., Wilson, J. D., McPhual, M. J. 1989. *Proc. Natl. Acad. Sci. USA* 86:327–31
99. Freedman, L. P., Luisi, B. F., Korszun, Z. R., Basavappa, R., Sigler, P. B., et al. 1988. *Nature* 334:543–46
100. Frankel, A. D., Pabo, C. O. 1988. *Cell* 53:675
101. Hard, T., Kellenbach, E., Boelens, R., Maler, B. A., Dahlman, K., et al. 1990. *Science* 249:157–60
102. Schwabe, J., Neuhaus, D., Rhodes, D. 1990. *Nature* 348:458–61
103. Luisi, B. F., Xu, W. X., Otwinowski, Z., Freedman, L. P., Yamamoto, K. R., et al. 1991. *Nature* 352:497–505
104. Landschulz, W. H., Johnson, P. F., McKnight, S. L. 1988. *Science* 240: 1759–64
105. Murre, C., McCaw, P. S., Baltimore, D. 1989. *Cell* 56:777–83
106. Hope, I. A., Struhl, K. 1986. *Cell* 46:885–94
107. Roman, C., Platero, J. S., Shuman, J., Calame, K. 1990. *Genes Dev.* 4:1404–15
108. Maekawa, T., Sakura, H., Kanei, I. C., Sudo, T., Yoshimura, T., et al. 1989. *EMBO J.* 8:2023–28
109. Hartings, H., Maddaloni, M., Lazzaroni, N., DiFonzo, N., Motto, M., et al. 1989. *EMBO J.* 8:2795–801
110. Fu, Y. H., Paietta, J. V., Mannix, D. G., Marzluf, G. A. 1989. *Mol. Cell. Biol.* 9:1120–27
111. O'Shea, E. K., Rutkowski, R., Kim, P. S. 1989. *Science* 243:538–42
112. O'Shea, E. K., Klemm, J. D., Kim, P. S., Alber, T. 1991. *Science* 254:539–44
113. Agre, P., Johnson, P. F., McKnight, S. L. 1989. *Science* 246:922–25

114. Talanian, R. V., McKnight, C. J., Kim, P. S. 1990. *Science* 249:769–71
115. O'Neil, K. T., Hoess, R. H., DeGrado, W. F. 1990. *Science* 249:774–78
116. Weiss, M. A., Ellenberger, T., Wobbe, C. R., Lee, J. P., Harrison, S. C., et al. 1990. *Nature* 347:575–78
117. Patel, L., Abate, C., Curran, T. 1990. *Nature* 347:572–74
118. Oakley, M. G., Dervan, P. B. 1990. *Science* 248:847–49
119. Vinson, C. R., Sigler, P. B., McKnight, S. L. 1989. *Science* 246:911–16
120. Nye, J. A., Graves, B. J. 1990. *Proc. Natl. Acad. Sci. USA* 87:3992–96
121. Rauscher, F. J. III, Cohen, D. R., Curran, T., Bos, T. J., Vogt, P. K., et al. 1988. *Science* 240:1010–16
122. Curran, T., Franza, B. R. J. 1988. *Cell* 55:395–97
123. Foulkes, N. S., Borrelli, E., Sassone-Corsi, P. 1991. *Cell* 64:739–49
124. Hai, T., Curran, T. 1991. *Proc. Natl. Acad. Sci. USA* 88:3720–24
125. Murre, C., McCaw, P. S., Vaessin, H., Caudy, M., Jan, L. Y., et al. 1989. *Cell* 58:537–44
126. Voronova, A., Baltimore, D. 1990. *Proc. Natl. Acad. Sci. USA* 87:4722–26
127. Prendergast, G. C., Ziff, E. B. 1989. *Nature* 341:392
128. Weintraub, H., Davis, R., Tapscott, S., Thayer, M., Krause, M., et al. 1991. *Science* 251:761–66
129. Benezra, R., Davis, R. L., Lockshon, D., Turner, D. L., Weintraub, H. 1990. *Cell* 61:49–59
130. Barinaga, M. 1991. *Science* 251:1176–77
131. Williams, T., Tjian, R. 1991. *Science* 251:1067–71
132. Breg, J. N., van Opheusden, J. H. J., Burgering, M. J. M., Boelens, R., Kaptein, R. 1990. *Nature* 346:586–89
133. Phillips, S. 1991. *Curr. Opin. Struc. Biol.* 1:89–98
134. Rafferty, J. B., Somers, W. S., Saint-Girons, I., Phillips, S. E. V. 1989. *Nature* 341:705–10
135. Vershon, A. K., Bowie, J. U., Karplus, T. M., Sauer, R. T. 1986. *Proteins* 1:302–11
136. Vershon, A. K., Liao, S-M., McClure, W. R., Sauer, R. T. 1987. *J. Mol. Biol.* 195:311–22
137. Brown, B. M., Bowie, J. U., Sauer, R. T. 1990. *Biochemistry* 29:11189–95
138. Youderian, P., Vershon, A., Bouvier, S., Sauer, R. T., Susskind, M. M. 1983. *Cell* 35:777–83
139. Knight, K. L., Sauer, R. T. 1989. *Proc. Natl. Acad. Sci. USA* 86:797–801
140. Knight, K. L., Sauer, R. T. 1989. *J. Biol. Chem.* 264:13706–10
141. Bowie, J. U., Sauer, R. T. 1990. *J. Mol. Biol.* 211:5–6
142. Yang, C-C., Nash, H. A. 1989. *Cell* 57:869–80
143. Tanaka, I., Appelt, K., Dijk, J., White, S. W., Wilson, K. S. 1984. *Nature* 310:374–81
144. White, S. W., Appelt, K., Wilson, K. S., Tanaka, I. 1989. *Proteins* 5:281–88
145. Pfeifer, K., Kim, K-S., Kogan, S., Guarente, L. 1989. *Cell* 56:291–301
146. Pan, T., Coleman, J. E. 1990. *Proc. Natl. Acad. Sci. USA* 87:2077–81
147. Green, L. M., Berg, J. M. 1989. *Proc. Natl. Acad. Sci. USA* 86:4047–51
148. Summers, M. F., South, T. L., Kim, B., Hare, D. R. 1990. *Biochemistry* 29:329–40
149. Joulin, V., Borics, D., Eleouet, J. F., Labastie, M-C., Chretien, S., et al. 1991. *EMBO J.* 10:1809–16
150. Tsai, S-F., Martin, D. I. K., Zon, L. I., D'Andrea, A. D., Wong, G. G., et al. 1989. *Nature* 339:446–51
151. Freemont, P. S., Hanson, I. M., Trowsdale, J. 1991. *Cell* 64:483–84
152. Freyd, G., Kim, S. K., Horvitz, H. R. 1990. *Nature* 344:876–79
153. Dameron, C. T., Winge, D. R., George, G. N., Sansone, M., Hu, S., et al. 1991. *Proc. Natl. Acad. Sci. USA* 88:6127–31
154. Karim, F. D., Urness, L. D., Thummel, C. S., Klemsz, M. J., McKercher, S. R., et al. 1991. *Genes Dev.* 4:1451–53
155. Travis, A., Amsterdam, A., Belanger, C., Grosschedl, R. 1991. *Genes Dev.* 5:880–94
156. Christ, C., Tye, B. 1991. *Genes Dev.* 5:751–63
157. Grotewold, E., Athma, P., Peterson, T. 1991. *Proc. Natl. Acad. Sci. USA* 88:4587–91
158. Bours, V., Villalobos, J., Burd, P. R., Kelly, K., Siebenlist, U. 1990. *Nature* 348:76–80
159. Ghosh, S., Gifford, A. M., Riviere, L. R., Tempst, P., Nolan, G. P., Baltimore, D. 1990. *Cell* 62:1019–29
160. Kieran, M., Blank, V., Logeat, F., Vandekerckhove, J., Lottspeich, F., et al. 1990. *Cell* 62:1007–18
161. Treisman, J., Harris, E., Desplan, C. 1991. *Genes Dev.* 5:594–604
162. Chalepakis, G., Fritsch, R., Fickenscher, H., Deutsch, U., Goulding, M., Gruss, P. 1991. *Cell* 66:873–84
163. Lai, E., Prezioso, V. R., Tao, W., Chen, W. S., Darnell, J. E. 1991. *Genes Dev.* 5:416–27

164. Weigel, D., Jackle, H. 1990. *Cell* 63:455–56
165. Burglin, F. 1991. *Cell* 66:11–12
166. Seeman, N. C., Rosenberg, J. M., Rich, A. 1976. *Proc. Natl. Acad. Sci. USA* 73:804–8
167. Pabo, C. O. 1984. *Specificity in protein-DNA interactions. Proc. Robert A. Welch Found. Conf. Chem. Res. XXVII, Stereospecificity in Chemistry and Biochemistry*, pp. 222–55
168. Dickerson, R. E., Drew, H. R. 1981. *J. Mol. Biol.* 149:761–86
169. Wemmer, D. E. 1991. *Curr. Opin. Struc. Biol.* 1:452–58
170. Shakked, Z. 1991. *Curr. Opin. Struc. Biol.* 1:446–51
171. Bowie, J. U., Luthy, R., Eisenberg, D. 1991. *Science* 253:164–70
172. Karplus, M., Petsko, G. A. 1990. *Nature* 347:631–39